An Interop Publication

ConneXions®

The Interoperability Report

October 1996 Volume 10, No. 10

ConneXions — The Interoperability Report tracks current and emerging standards and technologies within the computer and communications industry.

In this issue:

ConneXions is published monthly by Interop Company, a division of SOFTBANK Exposition and Conference Company, 303 Vintage Park Drive, Foster City, California, 94404–1138, USA.
Phone: +1 (415) 578-6900
Fax: +1 (415) 525-0194
E-mail: connexions@interop.com

Subscription hotline: 1-800-575-5717
or +1 610-892-1959

ISSN 0894–5926

From the Editor

It used to be said that you needed to learn UNIX in order to use the Internet. While this is no longer true, the success of the TCP/IP protocol suite is closely linked to its initial availability as part of the *Berkeley Software Distribution* (BSD) of UNIX. Additionally, UNIX has brought us a number of distributed computing concepts such as Remote Procedure Calls (RPC) and the Network Filesystem (NFS). In our "Back to Basics" series we take an in-depth look at NFS. The article is by Marshall Kirk McKusick, Keith Bostic, Michael J. Karels and John S. Quarterman, the authors of *The Design and Implementation of the 4.4BSD Operating System*. This book is highly recommended if you need to understand the internals of this popular networking software collection.

Multimedia communication is becoming more and more popular in the Internet, and new tools and protocols are under development. Most of the emerging applications use a form of communication in which a single sender transmits data to multiple receivers. This is called *multicast* communication. Our second article presents the basic principles of multicast services, discusses the problem of scalability with respect to group size and presents recent research approaches to overcoming existing bottlenecks.

SOFTBANK Expos recently announced the *Interop Graduate Institute* (IGI). The first educational program of its kind, it provides networking professionals with a comprehensive non-vendor specific understanding of the wide breadth of current technologies, and the knowledge to assess new technologies as they develop. The IGI, which will begin operation in December 1996, delivers a structured, university-level curriculum providing a rigorous education on computer networking technologies and underlying principles. With courses offered by premier networking experts, the IGI is designed to complement an engineering or computer science degree by providing conceptual and practical understanding of quickly evolving networking technologies. Graduates will receive Interop Graduate Institute certification, and the program will be reviewed by the American Council on Education for college accreditation status. The Institute program consists of five core courses which cover the fundamentals of networking, internetworking and client–server architecture, plus elective courses that cover advanced topics. Core courses include: Network Technologies, Network Interconnection and Internetworking, Network Protocols and Protocol Design, Routing and Routing Protocols, and Distributed Programming and Applications. The courses will be designed and approved by a curriculum committee comprised of top academic and industry experts. For more information, send e-mail to: ole@interop.com.

Back to Basics: The Network Filesystem (NFS)

by
Marshall Kirk McKusick, Keith Bostic, Michael J. Karels and John S. Quarterman

Introduction

This article is divided into three main sections. The first gives a brief history of remote filesystems. The second describes the client and server halves of NFS and the mechanics of how they operate. The final section describes the techniques needed to provide reasonable performance for remote filesystems in general, and NFS in particular.

History and Overview

When networking first became widely available in 4.2BSD, users who wanted to share files all had to log in across the net to a central machine on which the shared files were located. These central machines quickly became far more loaded than the user's local machine, so demand quickly grew for a convenient way to share files on several machines at once. The most easily understood sharing model is one that allows a server machine to export its filesystems to one or more client machines. The clients can then import these filesystems and present them to the user as though they were just another local filesystem.

Numerous remote-filesystem protocol designs and protocols were proposed and implemented. The implementations were attempted at all levels of the kernel. Remote access at the top of the kernel resulted in semantics that nearly matched the local filesystem, but had terrible performance. Remote access at the bottom of the kernel resulted in awful semantics, but great performance. Modern systems place the remote access in the middle of the kernel at the vnode layer. This level gives reasonable performance and acceptable semantics.

An early remote filesystem, *UNIX United,* was implemented near the top of the kernel at the system-call dispatch level. [23] It checked for file descriptors representing remote files and sent them off to the server. No caching was done on the client machine. The lack of caching resulted in slow performance, but in semantics nearly identical to a local filesystem. Because the current directory and executing files are referenced internally by vnodes rather than by descriptors, UNIX United did not allow users to change directory into a remote filesystem and could not execute files from a remote filesystem without first copying the files to a local filesystem.

At the opposite extreme was Sun Microsystem's *network disk,* implemented near the bottom of the kernel at the device-driver level. Here, the client's entire filesystem and buffering code was used. Just as in the local filesystem, recently read blocks from the disk were stored in the buffer cache. Only when a file access requested a block that was not already in the cache would the client send a request for the needed physical disk block to the server. The performance was excellent because the buffer cache serviced most of the file-access requests just as it does for the local filesystem. Unfortunately, the semantics suffered because of incoherency between the client and server caches. Changes made on the server would not be seen by the client, and vice versa. As a result, the network disk could be used only by a single client or as a read-only filesystem.

The first remote filesystem shipped with System V was *RFS* [16]. Although it had excellent UNIX semantics, its performance was poor, so it met with little use. Research at Carnegie-Mellon lead to the *Andrew filesystem* [4].

The Andrew filesystem was commercialized by Transarc and eventually became part of the Distributed Computing Environment promulgated by the Open Software Foundation, and was supported by many vendors. It is designed to handle widely distributed servers and clients and also to work well with mobile computers that operate while detached from the network for long periods.

NFS

The most commercially successful and widely available remote-filesystem protocol is the *Network Filesystem* (NFS) designed and implemented by Sun Microsystems [17, 21]. There are two important components to the success of NFS. First, Sun placed the protocol specification for NFS in the public domain. Second, Sun sells that implementation to all people who want it, for less than the cost of implementing it themselves. Thus, most vendors chose to buy the Sun implementation. They are willing to buy from Sun because they know that they can always legally write their own implementation if the price of the Sun implementation is raised to an unreasonable level. The 4.4BSD implementation was written from the protocol specification, rather than being incorporated from Sun, because of the developers desire to be able to redistribute it freely in source form.

NFS was designed as a client–server application. Its implementation is divided into a client part that imports filesystems from other machines and a server part that exports local filesystems to other machines. The general model is shown in Figure 1. Many goals went into the NFS design:

- The protocol is designed to be *stateless*. Because there is no state to maintain or recover, NFS can continue to operate even during periods of client or server failures. Thus, it is much more robust than a system that operates with state.
- NFS is designed to support UNIX filesystem semantics. However, its design also allows it to support the possibly less rich semantics of other filesystem types, such as MS-DOS.
- The protection and access controls follow the UNIX semantics of having the process present a UID and set of groups that are checked against the file's owner, group, and other access modes. The security check is done by filesystem-dependent code that can do more or fewer checks based on the capabilities of the filesystem that it is supporting. For example, the MS-DOS filesystem cannot implement the full UNIX security validation and makes access decisions solely based on the UID.
- The protocol design is transport independent. Although it was originally built using the UDP datagram protocol, it was easily moved to the TCP stream protocol. It has also been ported to run over numerous other non-IP-based protocols.

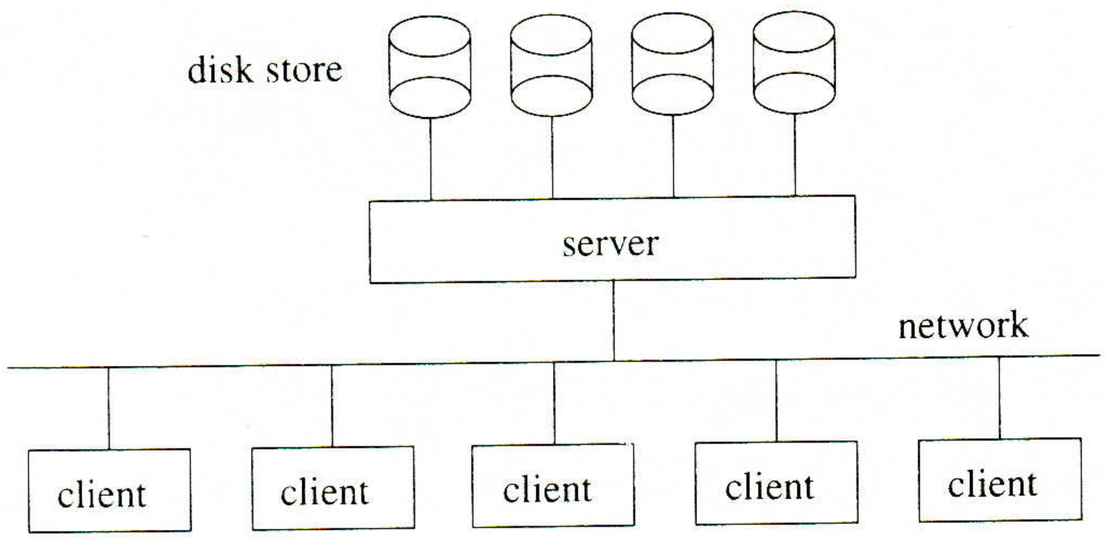

Figure 1: The division of NFS between client and server.

continued on next page

The Network Filesystem *(continued)*

Some of the design decisions limit the set of applications for which NFS is appropriate:

- The design envisions clients and servers being connected on a locally fast network. The NFS protocol does not work well over slow links or between clients and servers with intervening gateways. It also works poorly for mobile computing that has extended periods of disconnected operation.
- The caching model assumes that most files will not be shared. Performance suffers when files are heavily shared.
- The stateless protocol requires some loss of traditional UNIX semantics. Filesystem locking (*flock*) has to be implemented by a separate stateful daemon. Deferral of the release of space in an unlinked file until the final process has closed the file is approximated with a heuristic that sometimes fails.

Despite these limitations, NFS proliferated because it makes a reasonable tradeoff between semantics and performance; its low cost of adoption has now made it ubiquitous.

NFS structure and operation

NFS operates as a typical client–server application. The server receives *remote-procedure-call* (RPC) requests from its various clients. An RPC operates much like a local procedure call: The client makes a procedure call, then waits for the result while the procedure executes. For a remote procedure call, the parameters must be *marshalled* together into a message. Marshalling includes replacing pointers by the data to which they point and converting binary data to the canonical network byte order. The message is then sent to the server, where it is unmarshalled (separated out into its original pieces) and processed as a local filesystem operation. The result must be similarly marshalled and sent back to the client. The client splits up the result and returns that result to the calling process as though the result were being returned from a local procedure call [2]. The NFS protocol uses the Sun's RPC and external data-representation (XDR) protocols [15]. Although the kernel implementation is done by hand to get maximum performance, the user-level daemons described later in this section use Sun's public-domain RPC and XDR libraries.

The NFS protocol can run over any available stream- or datagram-oriented protocol. Common choices are the TCP stream protocol and the UDP datagram protocol. Each NFS RPC message may need to be broken into multiple packets to be sent across the network. A big performance problem for NFS running under UDP on an Ethernet is that the message may be broken into up to six packets; if any of these packets are lost, the entire message is lost and must be resent. When running under TCP on an Ethernet, the message may also be broken into up to six packets; however, individual lost packets, rather than the entire message, can be retransmitted. We discuss performance issues in greater detail later.

The set of RPC requests that a client can send to a server is shown in Table 1. After the server handles each request, it responds with the appropriate data, or with an error code explaining why the request could not be done. As noted in the table, most operations are *idempotent*. An idempotent operation is one that can be repeated several times without the final result being changed or an error being caused. For example, writing the same data to the same offset in a file is idempotent because it will yield the same result whether it is done once or many times.

However, trying to remove the same file more than once is non-idempotent because the file will no longer exist after the first try. Idempotency is an issue when the server is slow, or when an RPC acknowledgment is lost and the client retransmits the RPC request. The retransmitted RPC will cause the server to try to do the same operation twice. For a nonidempotent request, such as a request to remove a file, the retransmitted RPC, if undetected by the server recent-request cache [6], will cause a "no such file" error to be returned, because the file will have been removed already by the first RPC. The user may be confused by the error, because they will have successfully found and removed the file.

RPC request	Action	Idempotent
GETATTR	get file attributes	yes
SETATTR	set file attributes	yes
LOOKUP	look up file name	yes
READLINK	read from symbolic link	yes
READ	read from file	yes
WRITE	write to file	yes
CREATE	create file	yes
REMOVE	remove file	no
RENAME	rename file	no
LINK	create link to file	no
SYMLINK	create symbolic link	yes
MKDIR	create directory	no
RMDIR	remove directory	no
READDIR	read from directory	yes
STATFS	get filesystem attributes	yes

Table 1: NFS, Version 2, RPC requests.

Each file on the server can be identified by a unique *file handle*. A file handle is the token by which clients refer to files on a server. Handles are globally unique and are passed in operations, such as read and write, that reference a file. A file handle is created by the server when a path-name-translation request (lookup) is sent from a client to the server. The server must find the requested file or directory and ensure that the requesting user has access permission. If permission is granted, the server returns a file handle for the requested file to the client. The file handle identifies the file in future access requests by the client. Servers are free to build file handles from whatever information they find convenient. In the 4.4BSD NFS implementation, the file handle is built from a filesystem identifier, an inode number, and a *generation number*. The server creates a unique filesystem identifier for each of its locally mounted filesystems. A generation number is assigned to an inode each time that the latter is allocated to represent a new file. Each generation number is used only once. Most NFS implementations use a random-number generator to select a new generation number; the 4.4BSD implementation selects a generation number that is approximately equal to the creation time of the file. The purpose of the file handle is to provide the server with enough information to find the file in future requests. The filesystem identifier and inode provide a unique identifier for the inode to be accessed. The generation number verifies that the inode still references the same file that it referenced when the file was first accessed.

continued on next page

The Network Filesystem *(continued)*

The generation number detects when a file has been deleted, and a new file is later created using the same inode. Although the new file has the same filesystem identifier and inode number, it is a completely different file from the one that the previous file handle referenced. Since the generation number is included in the file handle, the generation number in a file handle for a previous use of the inode will not match the new generation number in the same inode. When an old-generation file handle is presented to the server by a client, the server refuses to accept it, and instead returns the "stale file handle" error message.

The use of the generation number ensures that the file handle is *time stable*. Distributed systems define a time-stable identifier as one that refers uniquely to some entity both while that entity exists and for a long time after it is deleted. A time-stable identifier allows a system to remember an identity across transient failures and allows the system to detect and report errors for attempts to access deleted entities.

The NFS protocol

The NFS protocol is *stateless*. Being stateless means that the server does not need to maintain any information about which clients it is serving or about the files that they currently have open. Every RPC request that is received by the server is completely self-contained. The server does not need any additional information beyond that contained in the RPC to fulfill the request. For example, a read request will include the credential of the user doing the request, the file handle on which the read is to be done, the offset in the file to begin the read, and the number of bytes to be read. This information allows the server to open the file, verifying that the user has permission to read it, to seek to the appropriate point, to read the desired contents, and to close the file. In practice, the server caches recently accessed file data. However, if there is enough activity to push the file out of the cache, the file handle provides the server with enough information to reopen the file.

In addition to reducing the work needed to service incoming requests, the server cache also detects retries of previously serviced requests. Occasionally, a UDP client will send a request that is processed by the server, but the acknowledgment returned by the server to the client is lost. Receiving no answer, the client will timeout and resend the request. The server will use its cache to recognize that the retransmitted request has already been serviced. Thus, the server will not repeat the operation, but will just resend the acknowledgment. To detect such retransmissions properly, the server cache needs to be large enough to keep track of at least the most recent few seconds of NFS requests.

The benefit of the stateless protocol is that there is no need to do state recovery after a client or server has crashed and rebooted, or after the network has been partitioned and reconnected. Because each RPC is self-contained, the server can simply begin servicing requests as soon as it begins running; it does not need to know which files its clients have open. Indeed, it does not even need to know which clients are currently using it as a server.

There are drawbacks to the stateless protocol. First, the semantics of the local filesystem imply state. When files are unlinked, they continue to be accessible until the last reference to them is closed. Because NFS knows neither which files are open on clients nor when those files are closed, it cannot properly know when to free file space.

As a result, it always frees the space at the time of the unlink of the last name to the file. Clients that want to preserve the freeing-on-last-close semantics convert unlink's of open files to renames to obscure names on the server. The names are of the form *.nfsAxxxx4.4,* where the *xxxx* is replaced with the hexadecimal value of the process identifier, and the A is successively incremented until an unused name is found. When the last close is done on the client, the client sends an unlink of the obscure filename to the server. This heuristic works for file access on only a single client; if one client has the file open and another client removes the file, the file will still disappear from the first client at the time of the remove. Other stateful semantics include advisory locking. The locking semantics cannot be handled by the NFS protocol. On most systems, they are handled by a separate lock manager; the 4.4BSD version of NFS does not implement them at all.

The second drawback of the stateless protocol is related to performance. For version 2 of the NFS protocol, all operations that modify the filesystem must be committed to stable-storage before the RPC can be acknowledged. Most servers do not have battery-backed memory; the stable store requirement means that all written data must be on the disk before they can reply to the RPC. For a growing file, an update may require up to three synchronous disk writes: one for the inode to update its size, one for the indirect block to add a new data pointer, and one for the new data themselves. Each synchronous write takes several milliseconds; this delay severely restricts the write throughput for any given client file.

Version 3 of the NFS protocol eliminates some of the synchronous writes by adding a new asynchronous write RPC request. When such a request is received by the server, it is permitted to acknowledge the RPC without writing the new data to stable storage. Typically, a client will do a series of asynchronous write requests followed by a commit RPC request when it reaches the end of the file or it runs out of buffer space to store the file. The commit RPC request causes the server to write any unwritten parts of the file to stable store before acknowledging the commit RPC. The server benefits by having to write the inode and indirect blocks for the file only once per batch of asynchronous writes, instead of on every write RPC request. The client benefits from having higher throughput for file writes. The client does have the added overhead of having to save copies of all asynchronously written buffers until a commit RPC is done, because the server may crash before having written one or more of the asynchronous buffers to stable store. When the client sends the commit RPC, the acknowledgment to that RPC tells which of the asynchronous blocks were written to stable store. If any of the asynchronous writes done by the client are missing, the client knows that the server has crashed during the asynchronous-writing period, and resends the unacknowledged blocks. Once all the asynchronously written blocks have been acknowledged, they can be dropped from the client cache.

The NFS protocol does not specify the granularity of the buffering that should be used when files are written. Most implementations of NFS buffer files in 8-Kbyte blocks. Thus, if an application writes 10 bytes in the middle of a block, the client reads the entire block from the server, modifies the requested 10 bytes, and then writes the entire block back to the server. The 4.4BSD implementation also uses 8-Kbyte buffers, but it keeps additional information that describes which bytes in the buffer are modified. If an application writes 10 bytes in the middle of a block, the client reads the entire block from the server, modifies the requested 10 bytes, but then writes back only the 10 modified bytes to the server.

continued on next page

The Network Filesystem *(continued)*

The block read is necessary to ensure that, if the application later reads back other unmodified parts of the block, it will get valid data. Writing back only the modified data has two benefits:

- Fewer data are sent over the network, reducing contention for a scarce resource.
- Nonoverlapping modifications to a file are not lost. If two different clients simultaneously modify different parts of the same file block, both modifications will show up in the file, since only the modified parts are sent to the server. When clients send back entire blocks to the server, changes made by the first client will be overwritten by data read before the first modification was made, and then will be written back by the second client.

The 4.4BSD NFS implementation

The NFS implementation that appears in 4.4BSD was written by Rick Macklem at the University of Guelph using the specifications of the Version 2 protocol published by Sun Microsystems [8, 19]. This NFS Version 2 implementation had several 4.4BSD-only extensions added to it; the extended version became known as the *Not Quite NFS* (NQNFS) protocol [9]. This protocol provides:

- Sixty-four-bit file offsets and sizes
- An access RPC that provides server permission checking on file open, rather than having the client guess whether the server will allow access
- An append option on the write RPC
- Extended file attributes to support 4.4BSD filesystem functionality more fully
- A variant of short-term *leases* with delayed-write client caching that give distributed cache consistency and improved performance [3].

Many of the NQNFS extensions were incorporated into the revised NFS Version 3 specification [14, 20]. Others, such as leases, are still available only with NQNFS. The NFS implementation distributed in 4.4BSD supports clients and servers running the NFS Version 2, NFS Version 3, or NQNFS protocol [10]. The NQNFS protocol is described later.

The 4.4BSD client and server implementations of NFS are kernel resident. NFS interfaces to the network with sockets using the kernel interface available through *sosend()* and *soreceive()*. There are connection-management routines for support of sockets using connection-oriented protocols; there are timeout and retransmit support for datagram sockets on the client side.

The less time-critical operations, such as mounting and unmounting, as well as determination of which filesystems may be exported and to what set of clients they may be exported are managed by user-level system daemons. For the server side to function, the *portmap, mountd,* and *nfsd* daemons must be running. The *portmap* daemon acts as a registration service for programs that provide RPC-based services. When an RPC daemon is started, it tells the *portmap* daemon to what port number it is listening and what RPC services it is prepared to serve. When a client wishes to make an RPC call to a given service, it will first contact the *portmap* daemon on the server machine to determine the port number to which RPC messages should be sent.

The interactions between the client and server daemons when a remote filesystem is mounted are shown in Figure 2. The *mountd* daemon handles two important functions:

- On startup and after a hangup signal, *mountd* reads the `/etc/exports` file and creates a list of hosts and passes this list into the kernel using the *mount* system call; the kernel links the list to the associated local filesystem mount structure so that the list is readily available for consultation when an NFS request is received.
- Client mount requests are directed to the *mountd* daemon. After verifying that the client has permission to mount the requested filesystem, *mountd* returns a file handle for the requested mount point. This file handle is used by the client for later traversal into the filesystem.

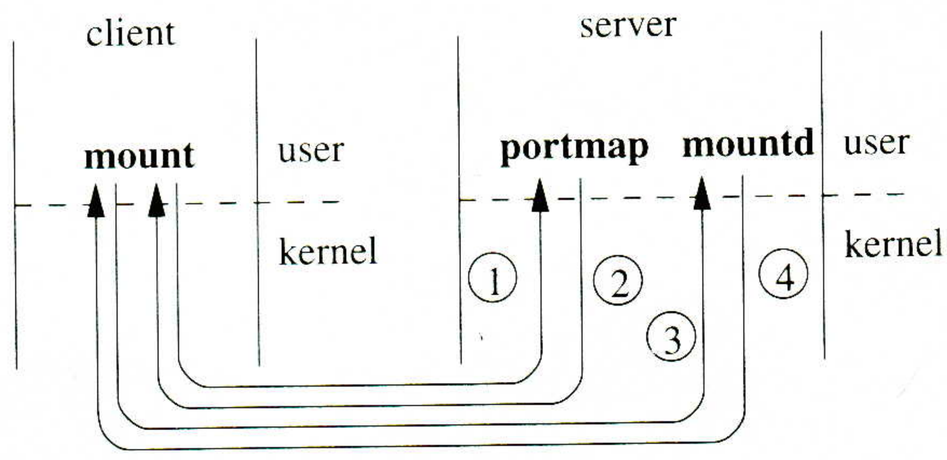

Figure 2: Daemon interaction when a remote filesystem is mounted

Step 1: The client's mount process sends a message to the well-known port of the server's *portmap* daemon, requesting the port address of the server's *mountd* daemon. Step 2: The server's *portmap* daemon returns the port address of its server's *mountd* daemon. Step 3: The client's *mount* process sends a request to the server's *mountd* daemon with the pathname of the filesystem that it wants to mount. Step 4: The server's *mountd* daemon requests a file handle for the desired mount point from its kernel. If the request is successful, the file handle is returned to the client's *mount* process. Otherwise, the error from the file-handle request is returned. If the request is successful, the client's *mount* process does a *mount* system call, passing in the file handle that it received from the server's *mountd* daemon.

The *nfsd* master daemon forks off children that enter the kernel using the *nfssvc* system call. The children normally remain kernel resident, providing a process context for the NFS RPC daemons. Typical systems run four to six *nfsd* daemons. If *nfsd* is providing datagram service, it will create a datagram socket when it is started. If *nfsd* is providing stream service, connected stream sockets will be passed in by the master *nfsd* daemon in response to connection-oriented connection requests from clients, When a request arrives on a datagram or stream socket, there is an upcall from the socket layer that invokes the *nfsrv_rcv()* routine. The *nfsrv_rcv()* call takes the message from the socket receive queue and dispatches that message to an available *nfsd* daemon. The *nfsd* daemon verifies the sender, and then passes the request to the appropriate local filesystem for processing. When the result returns from the filesystem, it is returned to the requesting client. The *nfsd* daemon is then ready to loop back and to service another request. The maximum degree of concurrency on the server is determined by the number of *nfsd* daemons that are started.

continued on next page

The Network Filesystem *(continued)*

For connection-oriented transport protocols, such as TCP, there is one connection for each client-to-server mount point. For datagram-oriented protocols, such as UDP, the server creates a fixed number of incoming RPC sockets when it starts its *nfsd* daemons; clients create one socket for each imported mount point. The socket for a mount point is created by the *mount* command on the client, which then uses it to communicate with the *mountd* daemon on the server. Once the client-to-server connection is established, the daemon processes on a connection-oriented protocol may do additional verification, such as Kerberos authentication. Once the connection is created and verified, the socket is passed into the kernel. If the connection breaks while the mount point is still active, the client will attempt a reconnect with a new socket.

The client side can operate without any daemons running, but the system administrator can improved performance by running several *nfsiod* daemons (these daemons provide the same service as the Sun *biod* daemons). The purpose of the *nfsiod* daemons is to do asynchronous read-aheads and write-behinds. They are typically started when the kernel begins running multiuser. They enter the kernel using the *nfssvc* system call, and they remain kernel resident, providing a process context for the NFS RPC client side. In their absence, each read or write of an NFS file that cannot be serviced from the local client cache must be done in the context of the requesting process. The process sleeps while the RPC is sent to the server, the RPC is handled by the server, and a reply sent back. No read-aheads are done, and write operations proceed at the disk-write speed of the server. When present, the *nfsiod* daemons provide a separate context in which to issue RPC requests to a server. When a file is written, the data are copied into the buffer cache on the client. The buffer is then passed to a waiting *nfsiod* that does the RPC to the server and awaits the reply. When the reply arrives, *nfsiod* updates the local buffer to mark that buffer as written. Meanwhile, the process that did the write can continue running. The Sun Microsystems reference port of the NFS protocol flushes all the blocks of a file to the server when that file is closed. If all the dirty blocks have been written to the server when a process closes a file that it has been writing, it will not have to wait for them to be flushed. The NQNFS protocol does not flush all the blocks of a file to the server when that file is closed.

When reading a file, the client first hands a read-ahead request to the *nfsiod* that does the RPC to the server. It then looks up the buffer that it has been requested to read. If the sought-after buffer is already in the cache because of a previous read-ahead request, then it can proceed without waiting. Otherwise, it must do an RPC to the server and wait for the reply. The interactions between the client and server daemons when I/O is done are shown in Figure 3.

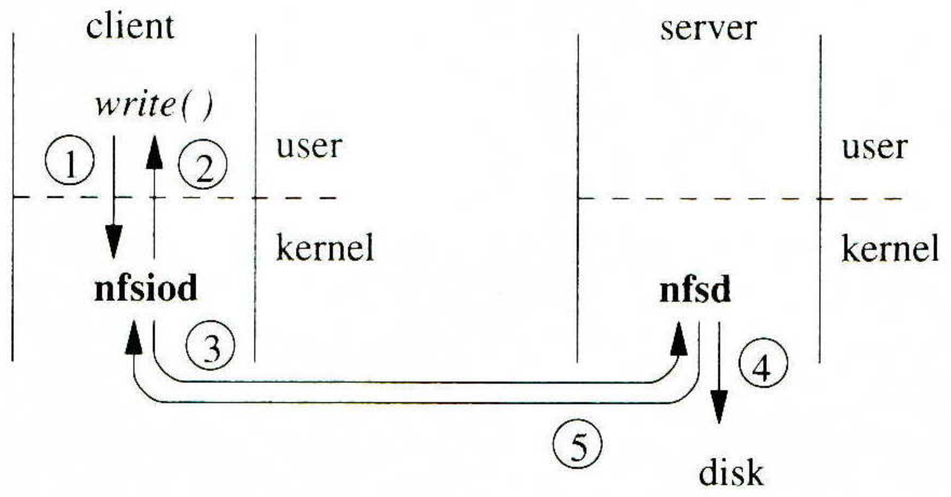

Figure 3: Daemon interaction when I/O is done.

Step 1: The client's process does a *write* system call. Step 2: The data to be written are copied into a kernel buffer on the client, and the *write* system call returns. Step 3: An *nfsiod* daemon awakens inside the client's kernel, picks up the dirty buffer, and sends the buffer to the server. Step 4: The incoming write request is delivered to the next available *nfsd* daemon running inside the kernel on the server. The server's *nfsd* daemon writes the data to the appropriate local disk, and waits for the disk I/O to complete. Step 5: After the I/O has completed, the server's *nfsd* daemon sends back an acknowledgment of the I/O to the waiting *nfsiod* daemon on the client. On receipt of the acknowledgment, the client's *nfsiod* daemon marks the buffer as clean.

Client–Server interactions

A local filesystem is unaffected by network service disruptions. It is always available to the users on the machine unless there is a catastrophic event, such as a disk or power failure. Since the entire machine hangs or crashes, the kernel does not need to concern itself with how to handle the processes that were accessing the filesystem. By contrast, the client end of a network filesystem must have ways to handle processes that are accessing remote files when the client is still running, but the server becomes unreachable or crashes. Each NFS mount point is provided with three alternatives for dealing with server unavailability:

- The default is a *hard mount* that will continue to try to contact the server "forever" to complete the filesystem access. This type of mount is appropriate when processes on the client that access files in the filesystem do not tolerate I/O system calls that return transient errors. A hard mount is used for processes for which access to the filesystem is critical for normal system operation. It is also useful if the client has a long-running program that simply wants to wait for the server to resume operation (e.g., after the server is taken down to run dumps).
- The other extreme is a *soft mount* that retries an RPC a specified number of times, and then the corresponding system call returns with a transient error. For a connection-oriented protocol, the actual RPC request is not retransmitted; instead, NFS depends on the protocol retransmission to do the retries. If a response is not returned within the specified time, the corresponding system call returns with a transient error. The problem with this type of mount is that most applications do not expect a transient error return from I/O system calls (since they never occur on a local filesystem). Often, they will mistakenly interpret the transient error as a permanent error, and will exit prematurely. An additional problem is deciding how long to set the timeout period. If it is set too low, error returns will start occurring whenever the NFS server is slow because of heavy load. Alternately, a large retry limit can result in a process hung for a long time because of a crashed server or network partitioning.
- Most system administrators take a middle ground by using an *interruptible mount* that will wait forever like a hard mount, but checks to see whether a termination signal is pending for any process that is waiting for a server response. If a signal (e.g., an interrupt) is sent to a process waiting for an NFS server, the corresponding I/O system call returns with a transient error. Normally, the process is terminated by the signal. If the process chooses to catch the signal, then it can decide how to handle the transient failure. This mount option allows interactive programs to be aborted when a server fails, while allowing long-running processes to await the server's return.

continued on next page

The Network Filesystem *(continued)*

The original NFS implementation had only the first two options. Since neither of these two options was ideal for interactive use of the filesystem, the third option was developed as a compromise solution.

RPC Transport issues

The NFS Version 2 protocol runs over UDP/IP transport by sending each request-reply message in a single UDP datagram. Since UDP does not guarantee datagram delivery, a timer is started, and if a timeout occurs before the corresponding RPC reply is received, the RPC request is retransmitted. At best, an extraneous RPC request retransmit increases the load on the server and can result in damaged files on the server or spurious errors being returned to the client when nonidempotent RPCs are redone. A recent-request cache normally is used on the server to minimize the negative effect of redoing a duplicate RPC request [6].

The amount of time that the client waits before resending an RPC request is called the *round-trip timeout* (RTT). Figuring out an appropriate value for the RTT is difficult. The RTT value is for the entire RPC operation, including transmitting the RPC message to the server, queuing at the server for an *nfsd,* doing any required I/O operations, and sending the RPC reply message back to the client. It can be highly variable for even a moderately loaded NFS server. As a result, the RTT interval must be a conservative (large) estimate to avoid extraneous RPC request retransmits. Adjusting the RTT interval dynamically and applying a congestion window on outstanding requests has been shown to be of some help with the retransmission problem [13].

On an Ethernet with the default 8-Kbyte read–write data size, the read–write reply-request will be an 8+-Kbyte UDP datagram that normally must be broken into at least six fragments at the IP layer for transmission. For IP fragments to be reassembled successfully into the IP datagram at the receive end, all fragments must be received at the destination. If even one fragment is lost or damaged in transit, the entire RPC message must be retransmitted, and the entire RPC redone. This problem can be exaggerated if the server is multiple hops away from the client through routers or slow links. It can also be nearly fatal if the network interface on the client or server cannot handle the reception of back-to-back network packets [7].

Using TCP

An alternative to all this madness is to run NFS over TCP transport, instead of over UDP. Since TCP provides reliable delivery with congestion control, it avoids the problems associated with UDP. Because the retransmissions are done at the TCP level, instead of at the RPC level, the only time that a duplicate RPC will be sent to the server is when the server crashes or there is an extended network partition that causes the TCP connection to break after an RPC has been received but not acknowledged to the client. Here, the client will resend the RPC after the server reboots, because it does not know that the RPC has been received.

The use of TCP also permits the use of read and write data sizes greater than the 8-Kbyte limit for UDP transport. Using large data sizes allows TCP to use the full duplex bandwidth of the network effectively, before being forced to stop and wait for RPC response from the server. NFS over TCP usually delivers comparable to significantly better performance than NFS over UDP, unless the client or server processor is slow. For processors running at less than 10 million instructions per second (MIPS), the extra CPU overhead of using TCP transport becomes significant.

The main problem with using TCP transport with Version 2 of NFS is that it is supported between only BSD clients and servers. However, the clear superiority demonstrated by the Version 2 BSD TCP implementation of NFS convinced the group at Sun Microsystems implementing NFS Version 3 to make TCP the default transport. Thus, a Version 3 Sun client will first try to connect using TCP; only if the server refuses will it fall back to using UDP.

Security issues

NFS is not secure because the protocol was not designed with security in mind. Despite several attempts to fix security problems, NFS security is still limited. Encryption is needed to build a secure protocol, but robust encryption cannot be exported from the United States. So, even if building a secure protocol were possible, doing so would be pointless, because all the file data are sent around the net in clear text. Even if someone is unable to get your server to send them a sensitive file, they can just wait until a legitimate user accesses it, and then can pick it up as it goes by on the net.

NFS export control is at the granularity of local filesystems. Associated with each local filesystem mount point is a list of the hosts to which that filesystem may be exported. A local filesystem may be exported to a specific host, to all hosts that match a subnet mask, or to all other hosts (the world). For each host or group of hosts, the filesystem can be exported read-only or read–write. In addition, a server may specify a set of subdirectories within the filesystem that may be mounted. However, this list of mount points is enforced by only the *mountd* daemon. If a malicious client wishes to do so, it can access any part of a filesystem that is exported to it.

The final determination of exportability is made by the list maintained in the kernel. So, even if a rogue client manages to snoop the net and to steal a file handle for the mount point of a valid client, the kernel will refuse to accept the file handle unless the client presenting that handle is on the kernel's export list. When NFS is running with TCP, the check is done once when the connection is established. When running with UDP, the check must be done for every RPC request.

The NFS server also permits limited remapping of user credentials. Typically, the credential for the superuser is not trusted and is remapped to the low-privilege user "nobody." The credentials of all other users can be accepted as given or also mapped to a default user (typically "nobody"). Use of the client UID and GID list unchanged on the server implies that the UID and GID space are common between the client and server (i.e., UID *N* on the client must refer to the same user on the server). The system administrator can support more complex UID and GID mappings by using the *umapfs* filesystem.

The system administrator can increase security by using Kerberos credentials, instead of accepting arbitrary user credentials sent without encryption by clients of unknown trustworthiness [18]. When a new user on a client wants to begin accessing files in an NFS filesystem that is exported using Kerberos, the client must provide a Kerberos ticket to authenticate the user on the server. If successful, the system looks up the Kerberos principal in the server's password and group databases to get a set of credentials, and passes in to the server *nfsd* a local translation of the client UID to these credentials. The *nfsd* daemons run entirely within the kernel except when a Kerberos ticket is received. To avoid putting all the Kerberos authentication into the kernel, the *nfsd* returns from the kernel temporarily to verify the ticket using the Kerberos libraries, and then returns to the kernel with the results.

continued on next page

The Network Filesystem *(continued)*

The NFS implementation with Kerberos uses encrypted timestamps to avert replay attempts. Each RPC request includes a timestamp that is encrypted by the client and decrypted by the server using a session key that has been exchanged as part of the initial Kerberos authentication. Each timestamp can be used only once, and must be within a few minutes of the current time recorded by the server. This implementation requires that the client and server clocks be kept within a few minutes of synchronization (this requirement is already imposed to run Kerberos). It also requires that the server keep copies of all timestamps that it has received that are within the time range that it will accept, so that it can verify that a timestamp is not being reused. Alternatively, the server can require that timestamps from each of its clients be monotonically increasing. However, this algorithm will cause RPC requests that arrive out of order to be rejected. The mechanism of using Kerberos for authentication of NFS requests is not well defined, and the 4.4BSD implementation has not been tested for interoperability with other vendors. Thus, Kerberos can be used only between 4.4BSD clients and servers.

Techniques for improving performance

Remote filesystems provide a challenging performance problem: Providing both a coherent networkwide view of the data and delivering that data quickly are often conflicting goals. The server can maintain coherency easily by keeping a single repository for the data and sending them out to each client when the clients need them; this approach tends to be slow, because every data access requires the client to wait for an RPC round-trip time. The delay is further aggravated by the huge load that it puts on a server that must service every I/O request from its clients. To increase performance and to reduce server load, remote filesystem protocols attempt to cache frequently used data on the clients themselves. If the cache is designed properly, the client will be able to satisfy many of the client's I/O requests directly from the cache. Doing such accesses is faster than communicating with the server, reducing latency on the client and load on the server and network. The hard part of client caching is keeping the caches coherent—that is, ensuring that each client quickly replaces any cached data that are modified by writes done on other clients. If a first client writes a file that is later read by a second client, the second client wants to see the data written by the first client, rather than the stale data that were in the file previously. There are two main ways that the stale data may be read accidentally:

- If the second client has stale data sitting in its cache, the client may use those data because it doesn't know that newer data are available.
- The first client may have new data sitting in its cache, but may not yet have written those data back to the server. Here, even if the second client asks the server for up-to-date data, the server may return the stale data because it does not know that one of its clients has a newer version of the file in that client's cache.

The second problem is related to the way that client writing is done. Synchronous writing requires that all writes be pushed through to the server during the *write* system call. This approach is the most consistent, because the server always has the most recently written data. It also permits any write errors, such as "filesystem out of space," to be sent back to the client process via the *write* call return. With an NFS filesystem using synchronous writing, error returns most closely parallel those from a local filesystem. Unfortunately, this approach restricts the client to only one write per RPC round-trip time.

An alternative to synchronous writing is delayed writing, where the *write* system call returns as soon as the data are cached on the client; the data are written to the server sometime later. This approach permits client writing to occur at the rate of local storage access up to the size of the local cache. Also, for cases where file truncation or deletion occurs shortly after writing, the write to the server may be avoided entirely, because the data have already been deleted. Avoiding the data push saves the client time and reduces load on the server.

There are some drawbacks to delayed writing. To provide full consistency, the server must notify the client when another client wants to read or write the file, so that the delayed writes can be written back to the server. There are also problems with the propagation of errors back to the client process that issued the *write* system call. For example, a semantic change is introduced by delayed-write caching when the file server is full. Here, delayed-write RPC requests can fail with an "out of space" error. If the data are sent back to the server when the file is closed, the error can be detected if the application checks the return value from the *close* system call. For delayed writes, written data may not be sent back to the server until after the process that did the write has exited—long after it can be notified of any errors. The only solution is to modify programs writing an important file to do an *fsync* system call and to check for an error return from that call, instead of depending on getting errors from *write* or *close*. Finally, there is a risk of the loss of recently written data if the client crashes before the data are written back to the server.

A compromise between synchronous writing and delayed writing is asynchronous writing. The write to the server is started during the *write* system call, but the *write* system call returns before the write completes. This approach minimizes the risk of data loss because of a client crash, but negates the possibility of reducing server write load by discarding writes when a file is truncated or deleted.

The simplest mechanism for maintaining full cache consistency is the one used by Sprite that disables all client caching of the file whenever concurrent write sharing might occur [12]. Since NFS has no way of knowing when write sharing might occur, it tries to bound the period of inconsistency by writing the data back when a file is closed. Files that are open for long periods are written back at 30-second intervals when the filesystem is synchronized. Thus, the NFS implementation does a mix of asynchronous and delayed writing, but always pushes all writes to the server on close. Pushing the delayed writes on close negates much of the performance advantage of delayed writing, because the delays that were avoided in the *write* system calls are observed in the *close* system call. With this approach, the server is always aware of all changes made by its clients with a maximum delay of 30 seconds and usually sooner, because most files are open only briefly for writing.

The server maintains read consistency by always having a client verify the contents of its cache before using that cache. When a client reads data, it first checks for the data in its cache. Each cache entry is stamped with an attribute that shows the most recent time that the server says that the data were modified. If the data are found in the cache, the client sends a timestamp RPC request to its server to find out when the data were last modified. If the modification time returned by the server matches that associated with the cache, the client uses the data in its cache; otherwise, it arranges to replace the data in its cache with the new data.

continued on next page

The Network Filesystem *(continued)*

The problem with checking with the server on every cache access is that the client still experiences an RPC round-trip delay for each file access, and the server is still inundated with RPC requests, although they are considerably quicker to handle than are full I/O operations. To reduce this client latency and server load, most NFS implementations track how recently the server has been asked about each cache block. The client then uses a tunable parameter that is typically set at a few seconds to delay asking the server about a cache block. If an I/O request finds a cache block and the server has been asked about the validity of that block within the delay period, the client does not ask the server again, but rather just uses the block. Because certain blocks are used many times in succession, the server will be asked about them only once, rather than on every access. For example, the directory block for the `/usr/include` directory will be accessed once for each `#include` in a source file that is being compiled. The drawback to this approach is that changes made by other clients may not be noticed for up to the delay number of seconds.

A more consistent approach used by some network filesystems is to use a *callback* scheme where the server keeps track of all the files that each of its clients has cached. When a cached file is modified, the server notifies the clients holding that file so that they can purge it from their cache. This algorithm dramatically reduces the number of queries from the client to the server, with the effect of decreasing client I/O latency and server load [4]. The drawback is that this approach introduces state into the server because the server must remember the clients that it is serving and the set of files that they have cached. If the server crashes, it must rebuild this state before it can begin running again. Rebuilding the server state is a significant problem when everything is running properly; it gets even more complicated and time consuming when it is aggravated by network partitions that prevent the server from communicating with some of its clients [11].

The 4.4BSD NFS implementation uses asynchronous writes while a file is open, but synchronously waits for all data to be written when the file is closed. This approach gains the speed benefit of writing asynchronously, yet ensures that any delayed errors will be reported no later than the point at which the file is closed. The implementation will query the server about the attributes of a file at most once every 3 seconds. This 3-second period reduces network traffic for files accessed frequently, yet ensures that any changes to a file are detected with no more than a 3-second delay. Although these heuristics provide tolerable semantics, they are noticeably imperfect. More consistent semantics at lower cost are available with the NQNFS lease protocol described in the next section.

Leases

The NQNFS protocol is designed to maintain full cache consistency between clients in a crash-tolerant manner. It is an adaptation of the NFS protocol such that the server supports both NFS and NQNFS clients while maintaining full consistency between the server and NQNFS clients. The protocol maintains cache consistency by using short-term leases instead of hard-state information about open files [3]. A *lease* is a ticket permitting an activity that is valid until some expiration time. As long as a client holds a valid lease, it knows that the server will give it a callback if the file status changes. Once the lease has expired, the client must contact the server if it wants to use the cached data.

Leases are issued using time intervals rather than absolute times to avoid the requirement of time-of-day clock synchronization. There are three important time constants known to the server. The *maximum_lease_term* sets an upper bound on lease duration—typically, 30 seconds to 1 minute. The *clock_skew* is added to all lease terms on the server to correct for differing clock speeds between the client and server. The *write_slack* is the number of seconds that the server is willing to wait for a client with an expired write-caching lease to push dirty writes.

Contacting the server after the lease has expired is similar to the NFS technique for reducing server load by checking the validity of data only every few seconds. The main difference is that the server tracks its clients' cached files, so there are never periods of time when the client is using stale data. Thus, the time used for leases can be considerably longer than the few seconds that clients are willing to tolerate possibly stale data. The effect of this longer lease time is to reduce the number of server calls almost to the level found in a full callback implementation such as the Andrew Filesystem [4]. Unlike the callback mechanism, state recovery with leases is trivial. The server needs only to wait for the lease's expiration time to pass, and then to resume operation. Once all the leases have expired, the clients will always communicate with the server before using any of their cached data. The lease expiration time is usually shorter than the time it takes most servers to reboot, so the server can effectively resume operation as soon as it is running. If the machine does manage to reboot more quickly than the lease expiration time, then it must wait until all leases have expired before resuming operation.

An additional benefit of using leases rather than hard state information is that leases use much less server memory. If each piece of state requires 64 bytes, a large server with hundreds of clients and a peak throughput of 2000 RPC requests per second will typically only use a few hundred Kbyte of memory for leases, with a worst case of about 3 Mbyte. Even if a server has exhausted lease storage, it can simply wait a few seconds for a lease to expire and free up a record. By contrast, a server with hard state must store records for all files currently open by all clients. The memory requirements are 3 to 12 Mbyte of memory per 100 clients served.

Whenever a client wishes to cache data for a file, it must hold a valid lease. There are three types of leases: non-caching, read caching, and write caching. A *noncaching lease* requires that all file operations be done synchronously with the server. A *read-caching lease* allows for client data caching, but no file modifications may be done. A *write-caching lease* allows for client caching of writes for the period of the lease. If a client has cached write data that are not yet written to the server when a write-cache lease has almost expired, it will attempt to extend the lease. If the extension fails, the client is required to push the written data.

If all the clients of a file are reading it, they will all be granted a read-caching lease. A read-caching lease allows one or more clients to cache data, but they may not make any modifications to the data. Figure 4 shows a typical read-caching scenario. The vertical solid black lines depict the lease records. Note that the time lines are not drawn to scale, since a client–server interaction will normally take less than 100 milliseconds, whereas the normal lease duration is 30 seconds. Every lease includes the time that the file was last modified. The client can use this timestamp to ensure that its cached data are still current.

continued on next page

The Network Filesystem *(continued)*

Initially, client A gets a read-caching lease for the file. Later, client A renews that lease and uses it to verify that the data in its cache are still valid. Concurrently, client B is able to obtain a read-caching lease for the same file.

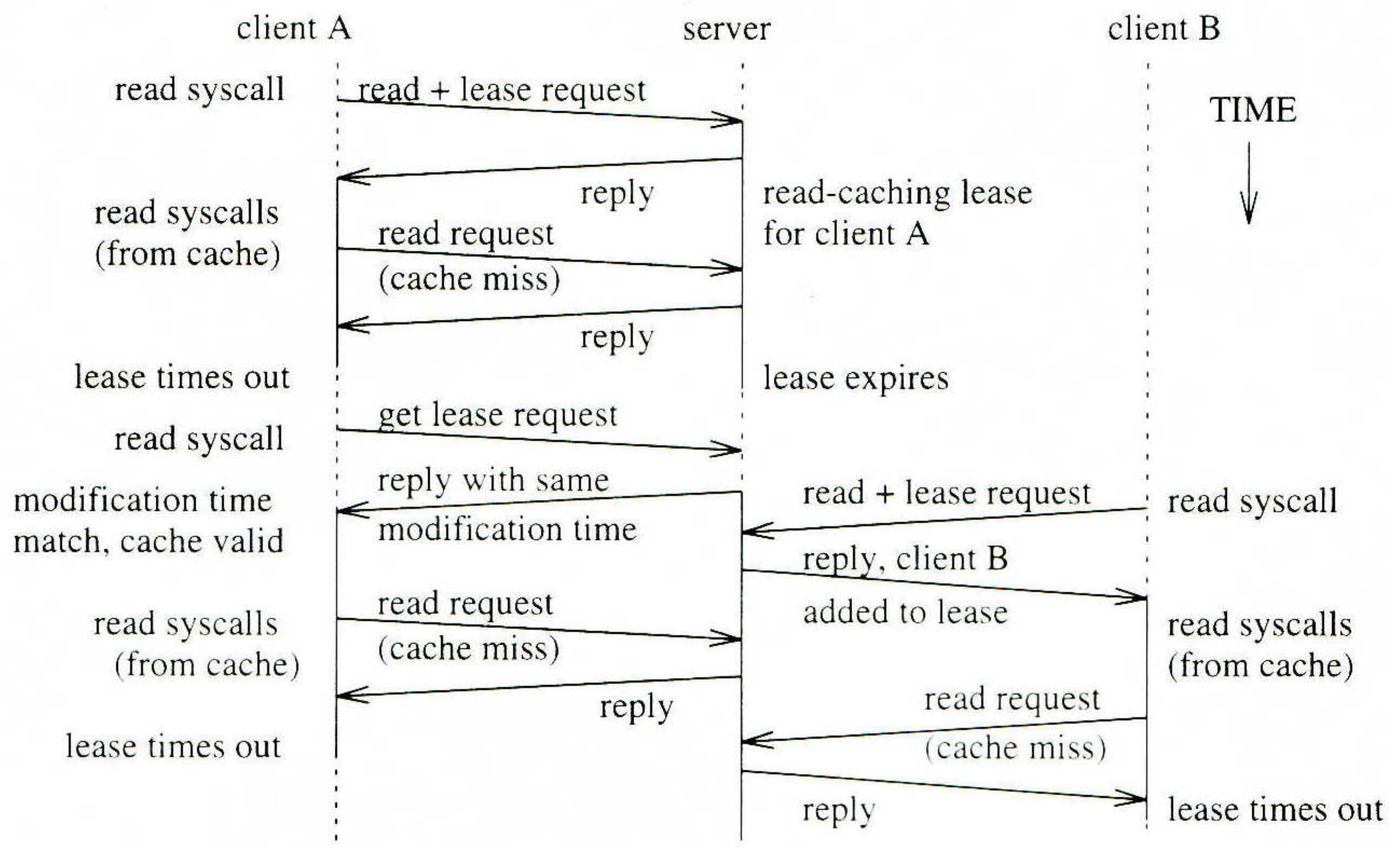

Figure 4: Read-caching leases.
Solid vertical lines represent valid leases.

If a single client wants to write a file and there are no readers of that file, the client will be issued a write-caching lease. A write-caching lease permits delayed write caching, but requires that all data be pushed to the server when the lease expires or is terminated by an *eviction notice*. When a write-caching lease has almost expired, the client will attempt to extend the lease if the file is still open, but is required to push the delayed writes to the server if renewal fails (see Figure 5). The writes may not arrive at the server until after the write lease has expired on the client. A consistency problem is avoided because the server keeps its write lease valid for *write_slack* seconds longer than the time given in the lease issued to the client. In addition, writes to the file by the lease-holding client cause the lease expiration time to be extended to at least *write_slack* seconds. This *write_slack* period is conservatively estimated as the extra time that the client will need to write back any written data that it has cached.

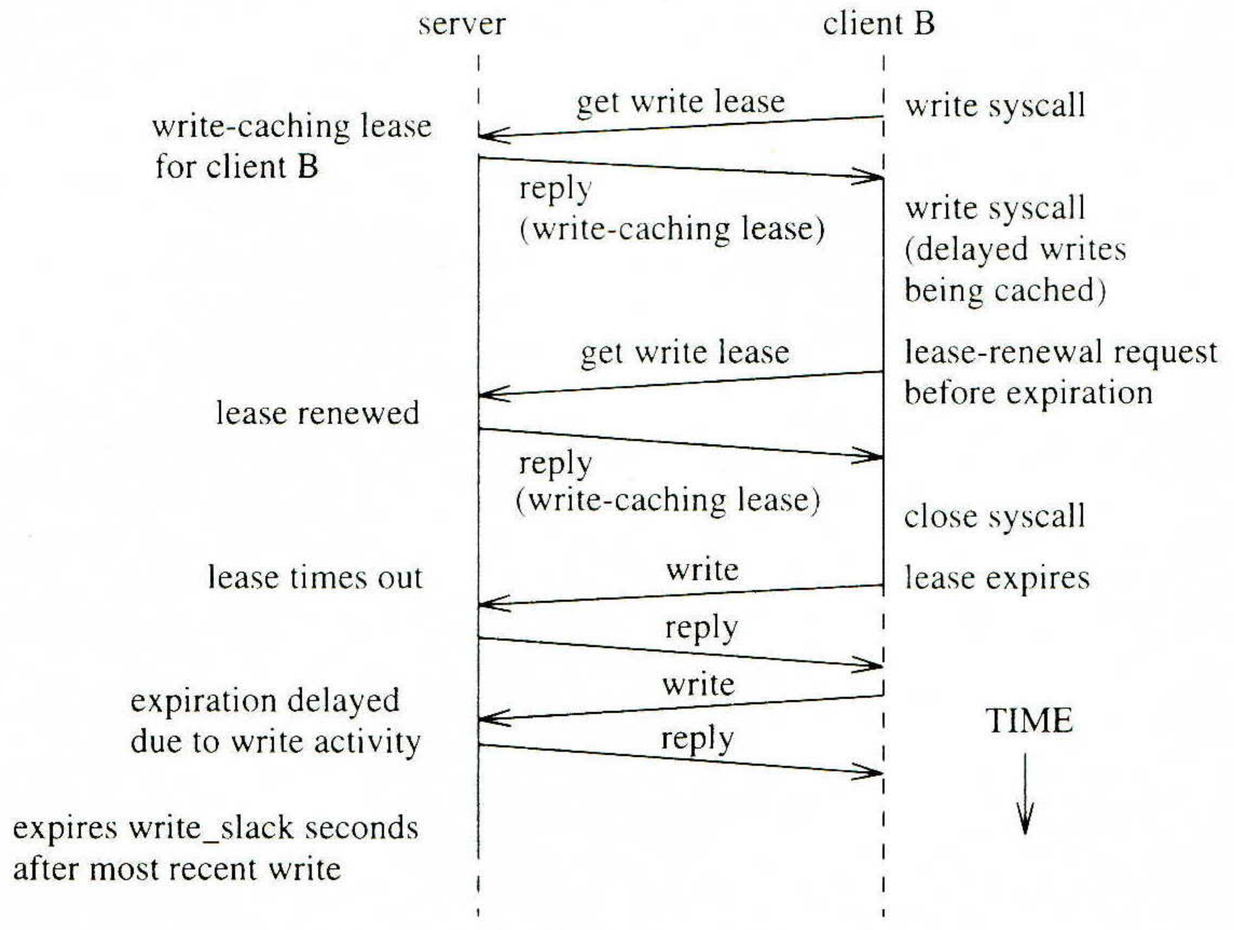

Figure 5: Write-caching lease.
Solid vertical lines represent valid leases.

If the value selected for *write_slack* is too short, a write RPC may arrive after the write lease has expired on the server. Although this write RPC will result in another client seeing an inconsistency, that inconsistency is no more problematic than the semantics that NFS normally provides.

The server is responsible for maintaining consistency among the NQNFS clients by disabling client caching whenever a server file operation would cause inconsistencies. The possibility of inconsistencies occurs whenever a client has a write-caching lease and any other client or a local operation on the server tries to access the file, or when a modify operation is attempted on a file being read cached by clients. If one of these conditions occurs, then all clients will be issued noncaching leases. With a noncaching lease, all reads and writes will be done through the server, so clients will always get the most recent data. Figure 6 shows how read and write leases are replaced by a noncaching lease when there is the potential for write sharing. Initially, the file is read by client A. Later, it is written by client B. While client B is still writing, client A issues another read request. Here, the server sends an "eviction notice" message to client B, and then waits for lease termination. Client B writes back its dirty data, then sends a "vacated" message. Finally, the server issues noncaching leases to both clients. In general, lease termination occurs when a "vacated" message has been received from all the clients that have signed the lease or when the lease has expired. The server does not wait for a reply for the message pair "eviction notice" and "vacated," as it does for all other RPC messages; they are sent asynchronously to avoid the server waiting indefinitely for a reply from a dead client.

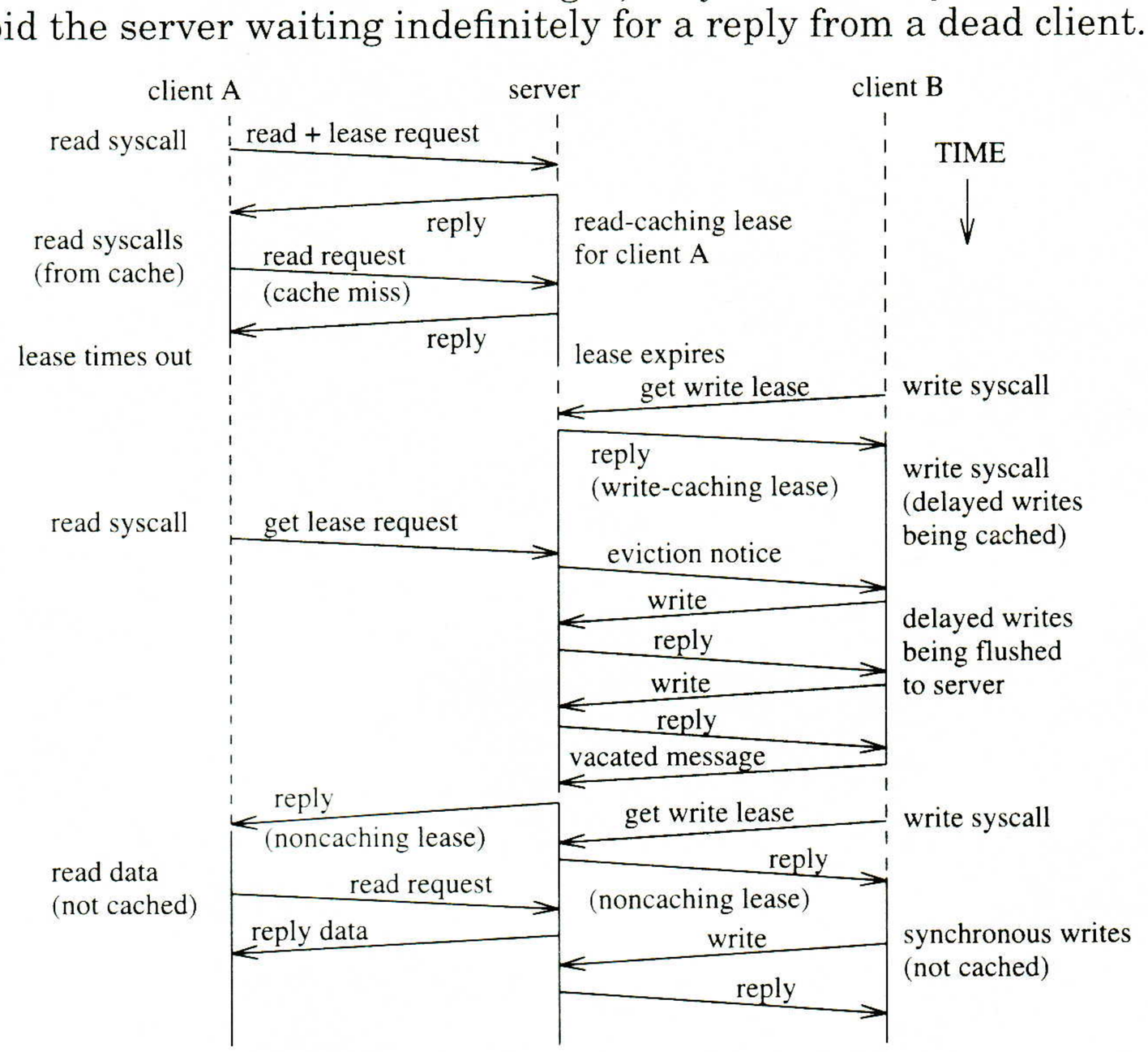

Figure 6: Write-sharing leases.
Solid vertical lines represent valid leases.

A client gets leases either by doing a specific lease RPC or by including a lease request with another RPC. Most NQNFS RPC requests allow a lease request to be added to them. Combining lease requests with other RPC requests minimizes the amount of extra network traffic.

continued on next page

The Network Filesystem *(continued)*

A typical combination can be done when a file is opened. The client must do an RPC to get the handle for the file to be opened. It can combine the lease request, because it knows at the time of the open whether it will need a read or a write lease. All leases are at the granularity of a file, because all NFS RPC requests operate on individual files, and NFS has no intrinsic notion of a file hierarchy. Directories, symbolic links, and file attributes may be read cached but are not write cached. The exception is the file-size attribute that is updated during cached writing on the client to reflect a growing file. Leases have the advantage that they are typically required only at times when other I/O operations occur. Thus, lease requests can almost always be piggy-backed on other RPC requests, avoiding some of the overhead associated with the explicit open and close RPC required by a long-term callback implementation.

The server handles operations from local processes and from remote clients that are not using the NQNFS protocol by issuing short-term leases for the duration of each file operation or RPC. For example, a request to create a new file will get a short-term write lease on the directory in which the file is being created. Before that write lease is issued, the server will vacate the read leases of all the NQNFS clients that have cached data for that directory. Because the server gets leases for all non-NQNFS activity, consistency is maintained between the server and NQNFS clients, even when local or NFS clients are modifying the filesystem. The NFS clients will continue to be no more or less consistent with the server than they were without leases.

Crash recovery

The server must maintain the state of all the current leases held by its clients. The benefit of using short-term leases is that, *maximum_lease_term* seconds after the server stops issuing leases, it knows that there are no current leases left. As such, server crash recovery does not require any state recovery. After rebooting, the server simply refuses to service any RPC requests except for writes (predominantly from clients that previously held write leases) until *write_slack* seconds after the final lease would have expired. For machines that cannot calculate the time that they crashed, the final-lease expiration time can be estimated safely as:

$$boot\_time + maximum\_lease\_term + write\_slack + clock\_skew$$

Here, *boot_time* is the time that the kernel began running after the kernel was booted. With a *maximum_lease_term* 30 to 60 seconds, and *clock_skew* and *write_slack* at most a few seconds, this delay amounts to about 1 minute, which for most systems is taken up with the server rebooting process. When this time has passed, the server will have no outstanding leases. The clients will have had at least *write_slack* seconds to get written data to the server, so the server should be up to date. After this, the server resumes normal operation.

There is another failure condition that can occur when the server is congested. In the worst-case scenario, the client pushes dirty writes to the server, but a large request queue on the server delays these writes for more than *write_slack* seconds. In an effort to minimize the effect of these *recovery storms*, the server replies "try again later" to the RPC requests that it is not yet ready to service [1]. The server takes two steps to ensure that all clients have been able to write back their written data. First, a write-caching lease is terminated on the server only when there are have been no writes to the file during the previous *write_slack* seconds.

Second, the server will not accept any requests other than writes until it has not been overloaded during the previous *write_slack* seconds. A server is considered overloaded when there are pending RPC requests and all its *nfsd* processes are busy.

Another problem that is solved by short-term leases is how to handle a crashed or partitioned client that holds a lease that the server wishes to vacate. The server detects this problem when it needs to vacate a lease so that it can issue a lease to a second client, and the first client holding the lease fails to respond to the vacate request. Here, the server can simply wait for the first client's lease to expire before issuing the new one to the second client. When the first client reboots or gets reconnected to the server, it simply reacquires any leases it now needs. If a client-to-server network connection is severed just before a write-caching lease expires, the client cannot push the dirty writes to the server. Other clients that can contact the server will continue to be able to access the file and will see the old data. Since the write-caching lease has expired on the client, the client will synchronize with the server as soon as the network connection has been re-established. This delay can be avoided with a write-through policy.

A detailed comparison of the effects of leases on performance is given in [9]. Briefly, leases are most helpful when a server or network is loaded heavily. Here, leases allow up to 30 to 50 percent more clients to use a network and server before beginning to experience a level of congestion equal to what they would on a network and server that were not using leases. In addition, leases provide better consistency and lower latency for clients, independent of the load. Although leases are new enough that they are not widely used in commercial implementations of NFS today, leases or a similar mechanism will need to be added to commercial versions of NFS if NFS is to be able to compete effectively against other remote filesystems, such as Andrew.

References

[1] M. Baker & J. Ousterhout, "Availability in the Sprite Distributed File System," ACM *Operating System Review,* Vol. 25, No. 2, p. 95–98, April 1991.

[2] A. D. Birrell & B. J. Nelson, "Implementing Remote Procedure Calls," ACM *Transactions on Computer Systems,*Vol. 2, No. 1, p. 39–59, February 1984.

[3] C. Gray & D. Cheriton, "Leases: An Efficient Fault-Tolerant Mechanism for Distributed File Cache Consistency," Proceedings of the Twelfth Symposium on Operating Systems Principles, p. 202–210, December 1989.

[4] J. Howard, "An Overview of the Andrew File System," USENIX Association Conference Proceedings, p. 23–26, January 1988.

[5] J. Howard, M. Kazar, S. Menees, D. Nichols, M. Satyanarayanan, R. Sidebotham, & M. West, "Scale and Performance in a Distributed File System," ACM *Transactions on Computer Systems,* Vol. 6, No. 1, p. 51–81, February 1988.

[6] C. Juszczak, "Improving the Performance and Correctness of an NFS Server," USENIX Association Conference Proceedings, p. 53–63, January 1989.

[7] C. Kent & J. Mogul, "Fragmentation Considered Harmful," Research Report 87/3, Digital Equipment Corporation Western Research Laboratory, December 1987.

continued on next page

The Network Filesystem *(continued)*

[8] R. Macklem, "Lessons Learned Tuning the 4.3BSD-Reno Implementation of the NFS Protocol," USENIX Association Conference Proceedings, p. 53–64, January 1991.

[9] R. Macklem, "Not Quite NFS, Soft Cache Consistency for NFS," USENIX Association Conference Proceedings, p. 261–278, January 1994.

[10] R. Macklem, "The 4.4BSD NFS Implementation," in *4.4BSD System Manager's Manual,* p. 6:1–14, O'Reilly & Associates, Inc., 1994.

[11] J. Mogul, "Recovery in Spritely NFS," Research Report 93/2, Digital Equipment Corporation Western Research Laboratory, June 1993.

[12] M. Nelson, B. Welch, & J. Ousterhout, "Caching in the Sprite Network File System," ACM *Transactions on Computer Systems,* Vol. 6, no. 1, p. 134–154, February 1988.

[13] B. Nowicki, "Transport Issues in the Network File System," *Computer Communications Review,* Vol. 19, No. 2, p. 16–20, April 1989.

[14] B. Pawlowski, C. Juszczak, P. Staubach, C. Smith, D. Lebel, & D. Hitz, "NFS Version 3: Design and Implementation," USENIX Association Conference Proceedings, p. 137–151, June 1994.

[15] Irving Reid, "RPCC: A Stub Compiler for Sun RPC," USENIX Association Conference Proceedings, p. 357–366, June 1987.

[16] A. Rifkin, M. Forbes, R. Hamilton, M. Sabrio, S. Shah, & K. Yueh, "RFS Architectural Overview," USENIX Association Conference Proceedings, p. 248–259, June 1986.

[17] R. Sandberg, D. Goldberg, S. Kleiman, D. Walsh, & B. Lyon, "Design and Implementation of the Sun Network Filesystem," USENIX Association Conference Proceedings, p. 119–130, June 1985.

[18] J. Steiner, C. Neuman, & J. Schiller, "Kerberos: An Authentication Service for Open Network Systems," USENIX Association Conference Proceedings, p. 191–202, February 1988.

[19] Sun Microsystems, "NFS: Network File System Protocol Specification," RFC 1094, March 1989.

[20] Sun Microsystems, "NFS: Network File System Version 3 Protocol Specification," June 1993.

[21] D. Walsh, B. Lyon, G. Sager, J. Chang, D. Goldberg, S. Kleiman, T. Lyon, R. Sandberg, & P. Weiss, "Overview of the Sun Network File System," USENIX Association Conference Proceedings, p. 117–124, January 1985.

[22] B. Clifford Neuman and Theodore Ts'o, "Kerberos: An Authentication Service for Computer Networks," *IEEE Communications,* Volume 32, No. 9, September 1994.

[23] David R. Brownbridge, Lindsay F. Marshall and Brian Randell, "The Newcastle Connection or UNIXes of the World Unite!" *Software Practice and Experience,* Volume 12, No. 12, December 1982, p. 1147–1162.

MARSHALL KIRK McKUSICK writes books and articles, consults, and teaches classes on UNIX- and BSD-related subjects. While at the University of California at Berkeley, he implemented the 4.2BSD fast file system, and was the Research Computer Scientist at the Berkeley Computer Systems Research Group (CSRG) overseeing the development and release of 4.3BSD and 4.4BSD. His particular areas of interest are the virtual-memory system and the filesystem. One day, he hopes to see them merged seamlessly. He earned his undergraduate degree in Electrical Engineering from Cornell University, and did his graduate work at the University of California at Berkeley, where he received Masters degrees in Computer Science and Business Administration, and a doctoral degree in Computer Science. He is a past president of the USENIX Association, and is a member of ACM and IEEE. In his spare time, he enjoys swimming, scuba diving, and wine collecting. The wine is stored in a specially constructed wine cellar (accessible from the net using the command "`telnet McKusick.com 451`") in the basement of the house that he shares with Eric Allman, his domestic partner of 17-and-some-odd years. E-mail: `mckusick@mckusick.com`

KEITH BOSTIC is a member of the technical staff at Berkeley Software Design, Inc. He spent 8 years as a member of the CSRG, overseeing the development of over 400 freely redistributable UNIX-compatible utilities, and is the recipient of the 1991 Distinguished Achievement Award from the University of California, Berkeley, for his work to make 4.4BSD freely redistributable. Concurrently, he was the principal architect of the 2.10BSD release of the Berkeley Software Distribution for PDP-11s, and the coauthor of the Berkeley Log Structured Filesystem and the Berkeley database package (DB). He is also the author of the widely used *vi* implementation, *nvi.* He received his undergraduate degree in Statistics and his Masters degree in Electrical Engineering from George Washington University. He is a member of the ACM, the IEEE, and several POSIX working groups. In his spare time, he enjoys scuba diving in the South Pacific, mountain biking, and working on a tunnel into Kirk and Eric's specially constructed wine cellar. He lives in Massachusetts with his wife, Margo Seltzer, and their cats. E-mail: `bostic@bsdi.com`

MICHAEL J. KARELS is the System Architect and Vice President of Engineering at Berkeley Software Design, Inc. He spent 8 years as the Principal Programmer of the CSRG at the University of California, Berkeley as the system architect for 4.3BSD. Karels received his Bachelor's degree in Microbiology from the University of Notre Dame. While a graduate student in Molecular Biology at the University of California, he was the principal developer of the 2.9BSD UNIX release of the Berkeley Software Distribution for the PDP-11. He is a member of the ACM, the IEEE, and several POSIX working groups. He lives with his wife Teri Karels in the backwoods of Minnesota. E-mail: `karels@bsdi.com`

JOHN S. QUARTERMAN is a partner in Texas Internet Consulting (TIC), which consults in networks and open systems with particular emphasis on TCP/IP networks, UNIX systems, and standards. He is the author of *The Matrix: Computer Networks and Conferencing Systems Worldwide* (Digital Press, 1990), and is a coauthor of *UNIX, POSIX, and Open Systems: The Open Standards Puzzle* (1993), *Practical Internetworking with TCP/IP and UNIX* (1993), *The Internet Connection: System Connectivity and Configuration* (1994), and *The E-Mail Companion: Communicating Effectively via the Internet and Other Global Networks* (1994), all published by Addison-Wesley. He is editor of *Matrix News,* a monthly newsletter about issues that cross network, geographic, and political boundaries, and of *Matrix Maps Quarterly;* both are published by Matrix Information and Directory Services, Inc. (MIDS) of Austin, Texas. He is a partner in Zilker Internet Park, which provides Internet access from Austin. He and his wife, Gretchen Quarterman, split their time among his home in Austin, hers in Buffalo, New York, and various other locations. E-mail: `jsq@tic.com`

[Ed.: This article is adapted from *The Design and Implementation of the 4.4BSD Operating System* by Marshall Kirk McKusick, Keith Bostic, Michael J. Karels, and John S. Quarterman, Addison-Wesley, 1996, ISBN 0-201-54979-4. Used with permission.]

Scalable Multicast Communication in the Internet

by Markus Hofmann, University of Karlsruhe

Introduction

Driven by the capabilities of modern high speed networks, a new generation of distributed systems is emerging. These systems make the support of superior distributed applications technically feasible, while increasing demand on communication in distributed environments has made them necessary. Forthcoming applications, such as collaborative distributed work, videoconferencing, or information dissemination, are expected to require information exchange between a large number of geographically dispersed components. They typically make use of a specific form of communication in which a single sender transmits data to multiple receivers. This form of communication is called *multicast* communication. This article presents the basic principles of efficiently providing multicast services, discusses the problem of scalability with respect to group size and presents recent research approaches to overcome existing bottlenecks.

Benefits of multicast

Multicast data transfer could be realized by repeatedly transmitting data units using point-to-point transfer to every communication participant. However, this approach does not scale well with the number of recipients and increases network load by the reciprocal of the group size. On the other hand, broadcasting data units may be an acceptable solution for small networks, but it causes a flood of data packets in wide-area networks. Rather than broadcasting information or using multiple unicast transfers, the forward-looking approach is to form a multicast group consisting of an arbitrary number of receivers and a single transmitter. Every data unit sent to the multicast group will be delivered only to those hosts that have registered themselves as members of the group. This requires special capabilities and additional mechanisms at different protocol levels.

Link level multicast

Protocol architectures designed to support multicast communication efficiently include special mechanisms even on the data link layer. Every data packet received by a network interface causes an interrupt and stimulates further processing at higher protocol levels. Therefore, it is desirable that hosts receive and process only those packets which are destined to them. Multicast protocols meet this desire by taking advantage of address filtering on the data link layer. Multicast capable interfaces can be configured to accept and process multicast packets in addition to directly addressed data frames. This renders possible the use of multicast addresses instead of the "all hosts" address or multiple unicast addresses. Transmitters connected to a broadcast media (e.g., Ethernet) need to send only one copy per packet without the consequence of causing further packet processing at non-member hosts.

Network level multicast

While transmitting one copy per packet is sufficient in local environments based on broadcast media, it is necessary to send multiple copies across ATM based networks or across heterogeneous internetworks (e.g., the Internet). In such a scenario, transfer of multicast messages can be optimized by delaying the replication of a data packet until it has to traverse different links. Therefore, routers and switches have to incorporate group management facilities as well as mechanisms to establish and maintain multicast routing trees. To enable early usage of multicast services without waiting for the availability of complete standards and without the need for wide-spread use of multicast-capable routers, the *Multicast Backbone* (Mbone) [1] has been established.

The Mbone is an overlay network on top of today's Internet providing a multicast facility to the network community. It makes use of Steve Deering's IP multicast extensions [2] and of multicast capable routing protocols, such as the *Distance Vector Multicast Routing Protocol* (DVMRP) [3] or the *Multicast Open Shortest Path First* (MOSPF) protocol [4]. However, research in the field of multicast routing is still ongoing [5]. New mechanisms are necessary to avoid explosion of states for wide-area routing and to support policy routing. A promising approach is the *Core Based Tree* (CBT) strategy [6], which reduces state information necessary to be kept within routers down to one per group.

Today's Mbone comprises only a small fraction of currently installed Internet routers and uses so called *tunnels* to link the multicast-capable islands together. These tunnels are manually configured by system administrators. They are used to forward multicast packets through non-multicast routers by encapsulating them inside regular IP packets. The Mbone has been established to get experience with the new multicast technology. However, the recent success of applications deployed over the Mbone illustrates the enormous potential of group communication and demonstrates the instant need for multicast services in wide-area networks.

Transport level multicast

Multicast capable networks provide efficient and scalable routing of data packets to multiple receivers. However, the bearer service provided by these communication networks does not fit the requirements of some applications. IP multicast, for example, offers an unreliable datagram service. The provision of reliable data transfer requires additional protocol mechanisms. According to the Internet protocol architecture, reliability as well as flow and rate control should be provided on an end-to-end basis. Therefore, mechanisms to support reliable multicast delivery should be integrated in transport level protocols.

When designing multicast protocols, scalability becomes more and more important. Widespread availability of IP multicast and development of applications deployed in the Mbone have considerably increased the geographic extent and the size of communication groups. Extensive use of services like Internet radio or large-scale conferencing leads to thousands of receivers being involved in a single multicast communication. In addition, communication participants may be spread all over the world. As the size and the geographic span of communication groups increases, efficient connection management schemes including scalable error and traffic control become more and more essential. Recent research projects deployed new transport level protocols to meet the requirements of large-scale multicast applications in heterogeneous networks. Some proposals only address particular aspects of group communication and focus only on some specific user environments. However, multimedia applications often have to handle several highly diverse data streams at the same time. A system supporting distributed cooperative work, for example, might offer shared use of a text editor while simultaneously providing audio and video communication between all the participants. Such applications require a multipurpose and flexible communication subsystem. On the other hand, other proposals suffer from a high degree of complexity and are too general in some issues. However, reliable data delivery is required by a wide range of applications. Therefore, this article will focus on protocol mechanisms for the provision of a reliable multicast service.

continued on next page

Multicast Communication in the Internet *(continued)*

Reliable multicast

Measurements have shown that packet losses in the current Mbone are significant. The data sets collected in [8] state that in one scenario almost 70% of transmitted packets were not successfully received by at least one receiver. This illustrates the need for efficient and powerful error correction schemes.

Some kind of interaction between sender and receivers is necessary to ensure correct data delivery as well as to perform any kind of congestion or flow control. Neither of the hosts has enough information to control data streams on its own. The provision of reliable data transfer, for example, is based on a comparison between sent and received data. The transmitter has knowledge about which data units have been sent and the receivers about which data units have been received successfully. Therefore, the provision of a reliable communication service requires the transmission of receiver status back to the sender or vice versa.

Sender-based error control

So called *sender-based* schemes, in which the transmitter is responsible for controlling data transfer, rely on collecting status information at the sending site. Receivers transmit control units, including acknowledgments and traffic control information, back to the sender. As the number of receivers becomes very large, the multicast sender is overwhelmed with return messages of its receivers. This condition is known as *sender implosion*. The effect of implosion is twofold. Firstly, the large number of return messages results in processing overhead at the sender and, therefore, delays data transfer. Secondly, an enormous amount of control units may cause an excess of both bandwidth and buffer space, which in turn causes additional message losses. While the latter issue is essential in wide-area networks and within large communication groups, processing overhead becomes more important in LAN/MAN environments. Over the last years, transmission capacity has been growing immensely, while protocol processing has turned out to be a performance bottleneck. Future photonic networks, for example, will provide bandwidth at no cost, but processing time will still remain a valuable resource. The processing bottleneck is particularly crucial for multicast communication. With an increasing number of receivers, processing of an increasing number of control packets needs to be performed. An optimal control procedure would reduce the number of status reports received by the sender down to one.

Receiver-based error control

Other approaches use *receiver-based* schemes, where receivers themselves are responsible for error detection and error recovery. They need not return status reports or acknowledgments to the transmitter. Instead, receivers use negative acknowledgments to request missed or corrupted data units from the transmitter. This reduces the number of return messages and prevents sender implosion. However, it is not possible to provide a full-reliable communication service while using such an error recovery scheme. The failure and dropping out of a single receiver could not be detected by the transmitter. Furthermore, flow or rate control always require some kind of status exchange between sender and receivers. This could lead to sender implosion even in the case of receiver-based reliable multicast communication.

Forward error correction

Forward error correction (FEC) has also been proposed to provide a reliable communication service. It is suitable for real-time applications since it allows error recovery without adding any delay associated with retransmissions. However, FEC does not prevent sender implosion caused by messages necessary for traffic control.

Moreover, adding redundant control information wastes bandwidth even when no or just a few errors occur. The sender must add enough control information to enable correction of all errors. While receivers in heterogeneous networks have highly diverse error characteristics, it is not adequate to choose a single fixed level of redundant coding. Such a fixed coding level may be excessive for some receivers while it may be insufficient for others. In addition, the error characteristic of receivers may change dynamically due to processing load, buffer occupancy, or network load. Therefore, the coding level should be adapted according to the current receiver state.

Retransmission schemes

Retransmission schemes may be quite effective, even for error control in multicast protocols geared to support real-time applications [9]. Retransmissions of data units are used to close gaps in the data stream of the receivers. Most protocols use Go-Back-N or selective repeat to retransmit lost and corrupted data. Receivers do request missed data units directly from the transmitter without any consideration of network topology and current network load. In the case of group communication, it is also possible to exchange data with neighboring receivers. It is preferable to request lost and corrupted data from a group member placed next to the host which is missing some information. An optimal error correction scheme would stimulate retransmissions of missed data units by the receiver located closest to the failing host. This would minimize transfer delay and network load. Studies of packet loss correlation in the current Mbone [8] show that packet loss is more likely to occur on the path between the multicast backbone and the local host rather than on the backbone links of the multicast tree. The measurements also show that, on average, there is just a little pair-wise spatially associated loss in the Mbone. Therefore, the probability that a receiver is able to get a missed data unit from a nearby group member is quite high.

Local Group Concept (LGC)

The described characteristics of the Mbone have strongly influenced the design and the development of a novel multicast protocol, which is called the *Local Group Concept* (LGC) [10]. The mechanisms of LGC are designed to support full-reliable and semi-reliable data transfer in large-scale, heterogeneous networks. They are based on a best-effort delivery model with multicast support. While these requirements are perfectly in conformity with IP and the current Mbone, the Local Group Concept is not restricted to the Internet protocol family. It can also be integrated in an extended ATM Adaptation Layer or in other protocol architectures with multicast support.

Defining Local Groups

The basic principle of LGC is to split the burden of acknowledgment handling and to distribute error correction among all the members of a multicast group. To achieve better scalability of point-to-multipoint services, LGC splits global communication groups into separate subgroups. These subgroups will combine communication participants within a local region, forming so called *Local Groups*. Each of them is represented by a *Local Group Controller* that collects status information from the members of its local group. The local Group Controller evaluates these return messages, combines them into a single control packet and transmits it back to the multicast sender or a higher level local Group Controller. Local Group Controllers also support the provision of local retransmissions. They coordinate local recovery from data loss to avoid expensive retransmissions from the multicast sender. This reduces delay and decreases the load for transmitter and network. The integration of message processing capabilities into local Group Controllers reduces the implosion problem of multicast traffic and error control for large groups.

continued on next page

Multicast Communication in the Internet *(continued)*

Local Group Controllers evaluate received control units and inform the multicast sender about the status of the Local Group. This includes error reports as well as parameters to control data flow. Parallel processing of status reports and their combination into a single message per Local Group relieves the multicast sender as it reduces the number of control units to be evaluated at the sending side.

In each local group, one of the receivers is determined to function as local Group Controller. The dedicated system has to collect control messages from all the members of its subgroup and has to forward them to the multicast sender in a single composite control unit. Controllers of subgroups are also responsible for organizing local retransmissions. After evaluating received status messages a local Group Controller tries to transfer lost data units to all the receivers that have observed errors or losses. To retransmit data units a local Group Controller can use either unicast or restricted multicast transmission. This decision may be static or dynamically based on the number of failed receivers. If a controller itself misses a data unit, it will ask another group member to multicast this data unit to the local group. Therefore, a multicast sender has only to retransmit messages missed by all members of a subgroup. Local retransmissions lead to shorter delays and decrease the number of data units flowing through a global network.

Example

An example scenario illustrating the basic idea and the advantage of this concept is given in Figure 1. A multicast sender communicates over a satellite link with four receivers, which are connected to a common switch. The satellite link is characterized by high transfer delay and high carrier fees. Therefore, it is desirable to reduce data traffic over this link. In this type of scenario it is useful to combine all four receivers into a single subgroup. One of the receiving hosts has to function as the controller of the subgroup. In this case, local retransmissions do not traverse the satellite link. This reduces transfer delay and network load within the satellite link.

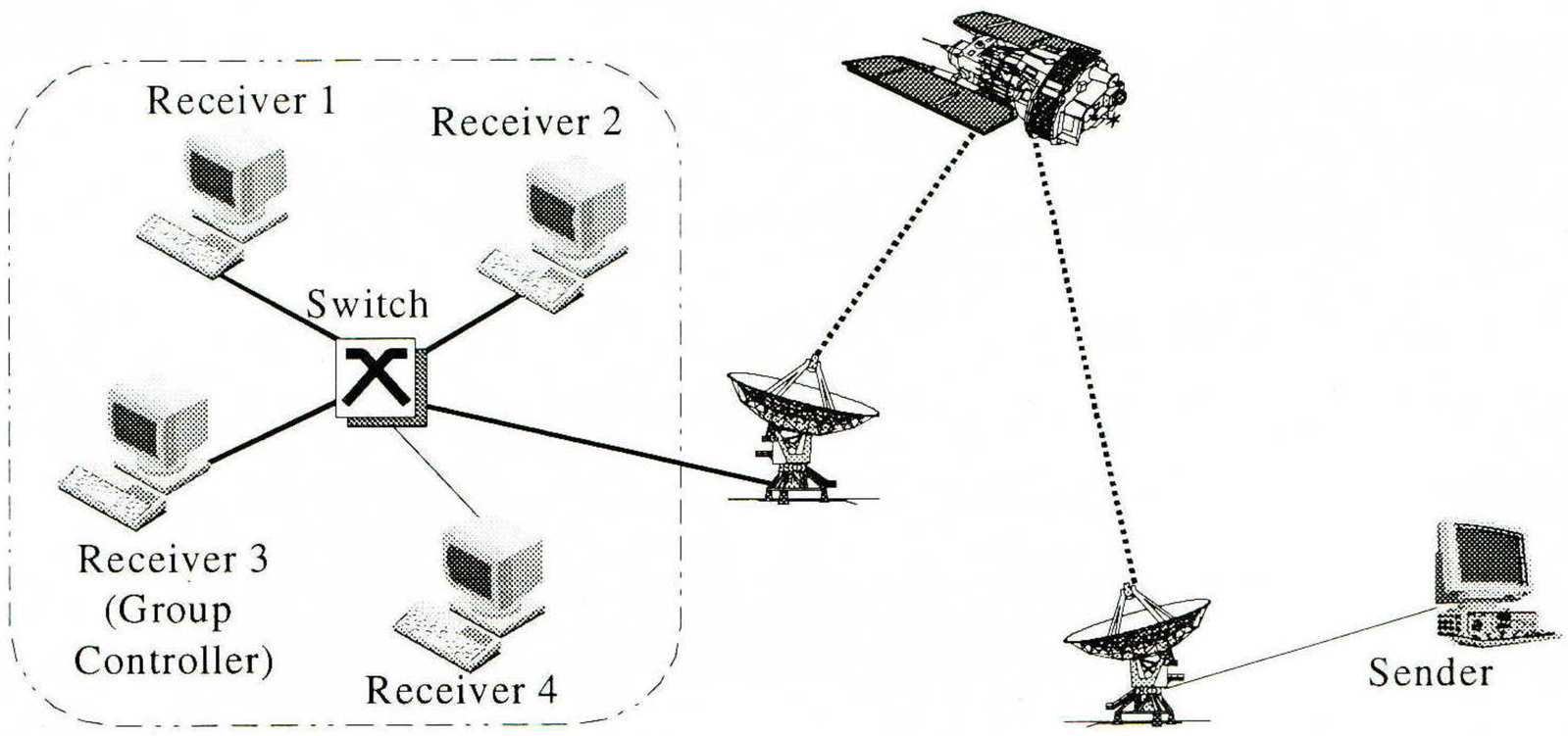

Figure 1: Example for the Definition of Local Groups

The resulting data streams are shown in Figure 2. The transmitter multicasts data units directly to all group members using a multicast capable delivery service (1). The local Group Controllers are kept out of the outgoing data path avoiding an extra handling of data units at each level of the hierarchy. Therefore, local Group Controllers need not to store data fragments, reassemble complete data units, interpret and forward them. Instead, multicast forwarding is done by the delivery service in a more efficient way.

After receiving a status request, regular receivers transmit control messages to their corresponding local Group Controller (2). The controller combines the status reports into a single control unit and sends it to the multicast transmitter (3). Therefore, it could be said that the Local Group Concept causes some kind of triangulation of data flow.

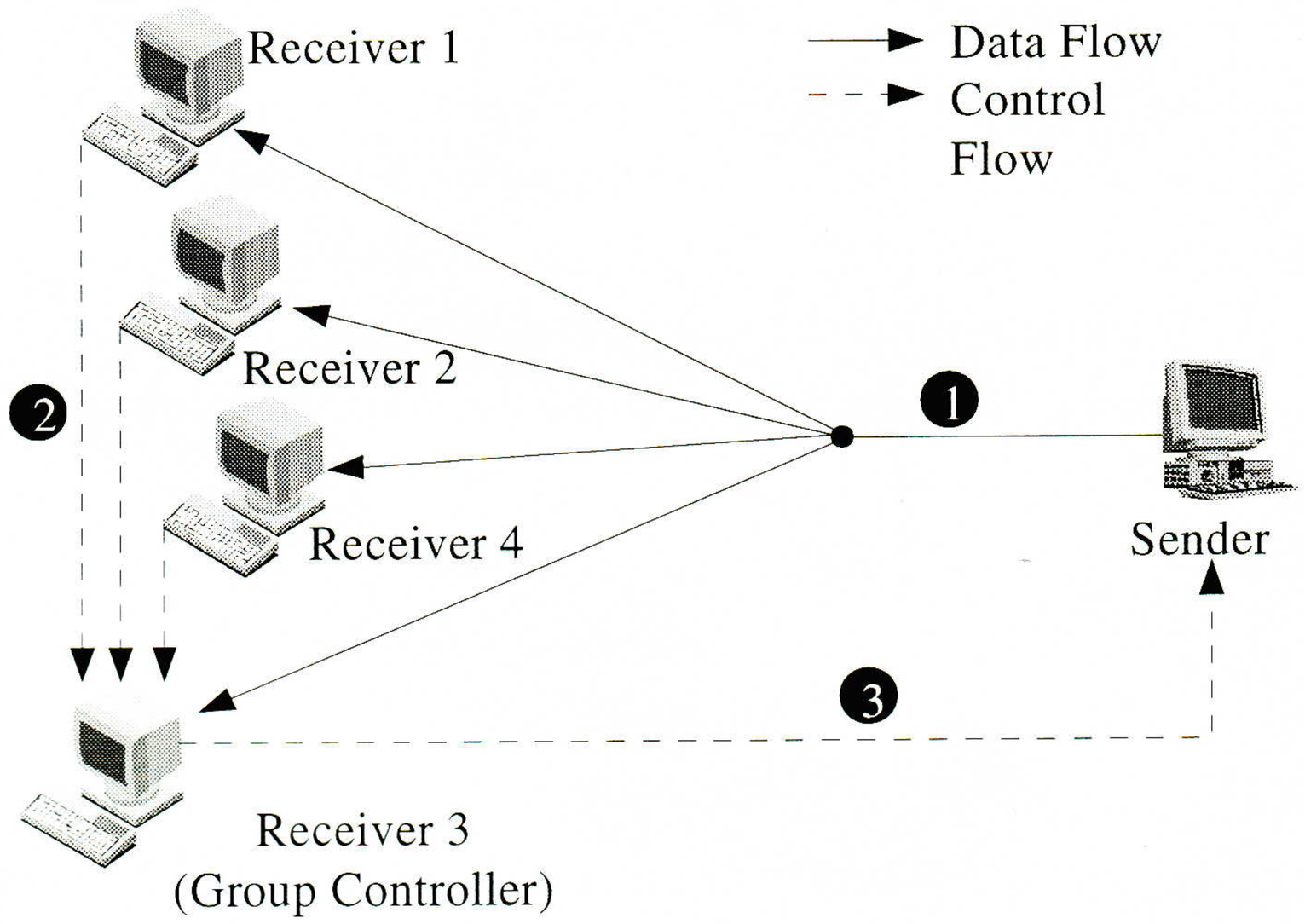

Figure 2: Triangulation of Data Flow

Local Group reformation

The Local Group Concept also includes mechanisms to increase fault tolerance [10]. The Fail Stop of a local Group Controller, for example, is handled by dynamic reformation of Local Groups. Of course, actions performed after the failure of a communication participant depend on the given group semantic. If the communication user requires an all-reliable service, the multicast connection will be closed due to the failure of a group member. In the case of semi-reliable service, the actions to be performed depend on the attributes and the role of the failed receiver. Dynamic group reformation is also used to adapt the group structure to the current network load and to the current state of all communication participants. For further improvements in respect to transfer delay and implosion, Local Groups are organized in a hierarchical structure.

Metrics

In the above example, the decision to combine all four receivers into one single subgroup has been based on the intention of minimizing transfer delay. In other scenarios, it may be appropriate to minimize other parameters. The suitability of a certain metric, such as delay, bandwidth, throughput, error probability, reliability, carrier fees, or number of hops between two nodes, depends mainly on the application using a communication service. While an interactive application may wish to minimize transfer delay, a user transferring files is interested in reducing the financial cost of a transfer. Of course, it could also be suitable to combine several metrics and to weight them according to the intention of the service user. The establishment of Local Groups is also influenced by the structure and the type of a network used for data exchange. At the moment, a new communication service is under development to support application-specific and automated establishment of a group hierarchy.

continued on next page

Multicast Communication in the Internet *(continued)*

Related work

Several other approaches have been proposed to provide a reliable multicast service. Early multicast protocols have used sender-based error control to achieve reliability. However, this is not suitable for large-scale group communication. Recent approaches prefer receiver-based control or hybrid schemes and include special techniques to prevent sender implosion.

Xpress Transport Protocol

The first protocol to incorporate mechanisms for implosion control has been the *Xpress Transport Protocol* (XTP) [11]. It has defined two heuristics called *damping* and *slotting*. These algorithms suppress redundant control messages by multicasting return messages to the whole group. Hence, every group member receives status messages of other members and skips its own status report if the incoming control unit corresponds to its own state. This mechanism reduces the number of control messages to be processed by the sender. Nevertheless, multicasting control units may be inefficient in large-scale, wide-area networks.

Scalable Reliable Multicast

The *Scalable Reliable Multicast* (SRM) [12] enhances damping and slotting mechanisms of XTP to reduce state management overhead. Receivers take solely the responsibility for error correction which is why SRM achieves a high degree of fault tolerance. However, a transmitter is not able to detect the failure of a single receiver. The protocol has been designed for use in the whiteboard tool *wb* [13]. SRM is an example of the receiver-based approach for error control. A receiver missing a certain data unit multicasts a *repair request* to the whole group. Group members that have successfully received the requested packet will multicast it to the entire group. To avoid a flood of repair requests and of retransmission, SRM suppress redundant requests by using timers carefully set and adjusted to the current network load. The efficiency of the protocol mainly depends on the correct setting of these timers.

Multicast Transport Protocol

The *Multicast Transport Protocol* (MTP) [14] realizes a centralized control scheme to provide a reliable, totally ordered multicast delivery. Data units from multiple transmitters are delivered in the same order to all group members. A so-called *master* controls data flow by assigning tokens for data transmission. Each potential sender has to obtain a token from the master before transmitting data to the group. This mechanism maintains the global order of data units. Error recovery in MTP is based on negative acknowledgments and retransmissions by the data source.

Reliable Multicast Protocol

The *Reliable Multicast Protocol* (RMP) [15] runs on top of IP multicast and provides a reliable, totally ordered, atomic multicast delivery. It is based on negative acknowledgments that are multicasted to avoid implosion. Reliability is ensured by a rotating token scheme. A single token is passed between group members and designates the site to multicast an acknowledgment for the recently received packets. Missed data units are retransmitted via multicast to all group members.

Other approaches

The *Reliable Multicast Transport Protocol* (RMTP) [16] and the *Tree-based Multicast Transport Protocol* (TMTP) [17] both use a hierarchical group structure similar to LGC. In contrast to the Local Group Concept, both approaches do not make use of data exchange between neighboring receivers. Instead, both approaches follow strong hierarchical guidelines and request missed data units always at a higher level controller.

Summary/Conclusion

Common multicast protocols are a significant improvement over simple point-to-point protocols. However, most of them are not suitable for the case where the transmitter has to handle data flow to an large number of receivers. To avoid sender implosion and to increase efficiency of error recovery, the Local Group Concept has been introduced. The benefits are achieved without the necessity to modify internal network equipment such as ATM switches or IP routers. Work is ongoing to develop a new service for automated establishment of group hierarchies.

Acknowledgments

The author would like to thank J. William Atwood from Concordia University Montreal for valuable discussions. Special thanks to all the members of the LGC Working Group at University of Karlsruhe for valuable comments and suggestions on various subjects of this article.

Author's address

Markus Hofmann
Institute of Telematics
University of Karlsruhe
Zirkel 2
D-76128 Karlsruhe
GERMANY
Fax: +49 721 608 3982
E-Mail: m.hofmann@ieee.org
Web: http://www.telematik.informatik.uni-karlsruhe.de/~hofmann

References

[1] V. Kumar, *Mbone: Interactive Multimedia on the Internet,* New Riders Publishing, Indianapolis, Indiana, USA, 1995.

[2] S. Deering, "Host Extensions for IP Multicasting," RFC 1112, August 1989.

[3] S. Deering, C. Partridge, D. Waitzman, "Distance Vector Multicast Routing Protocol," RFC 1075, November 1988.

[4] J. Moy, "Multicast Extensions to OSPF," RFC 1584, March 1994.

[5] C. Huitema, *Routing in the Internet,* ISBN 0-13-132192-7, Prentice Hall, New Jersey, 1995.

[6] A. Ballardie, P. Francis, J. Crowcroft, "Core Based Tree (CBT), An Architecture for Scalable Inter-Domain Routing," ACM-SIGCOM '93, September 1993, pp. 85–95.

[7] B. Carpenter, "Architectural Principles of the Internet," RFC 1958, June 1996.

[8] M. Yajnik, J. Kurose, D. Towsley, "Packet Loss Correlation in the Mbone Multicast Network," UMASS CMPSCI Technical Report # 96-32, University of Massachusetts at Amherst, 1995.

[9] S. Pejhan, M. Schwartz, D. Anastassiou, "Error Control Using Retransmission Schemes in Multicast Transport Protocols for Real-Time Media," IEEE/ACM *Transactions on Networking,* Volume 4, No. 3, pp. 413–427, June 1996.

[10] M. Hofmann, "A Generic Concept for Large-Scale Multicast," Proceedings of International Zurich Seminar on Digital Communications, Springer Verlag, February 1996.

[11] W. T. Strayer, ed., "Xpress Transport Protocol Specification, Revision 4.0," Available from XTP Forum, Santa Barbara, USA, March 1995.

continued on next page

Multicast Communication in the Internet *(continued)*

[12] S. Floyd, V. Jacobson, S. McCanne, C. Liu, L. Zhang, "A Reliable Multicast Framework for Light-weight Sessions and Application Level Framing," *Computer Communication Review,* Volume 25, No. 4, Proceedings of ACM SIGCOMM '95, August 1995.

[13] V. Jacobson, "A Portable, Public Domain Network Whiteboard," Xerox Parc, viewgraphs, April 1992.

[14] S. Armstrong, A. Freier, K. Marzullo, "Multicast Transport Protocol," RFC 1301, February 1992.

[15] B. Whetten. T. Montgomery, S. Kaplan, "A High Performance Totally Ordered Multicast Protocol," Submitted to INFOCOM '95, April 1995.

[16] J. C. Lin, S. Paul, "RMTP: A reliable Multicast Transport Protocol," IEEE INFOCOM '96, 1996.

[17] R. Yavatkar, J. Griffioen, M. Sudan, "A Reliable Protocol for Interactive Collaborative Applications," ACM Multimedia '95, 1995.

[18] R. Braden & L. Zhang, "RSVP: A Resource ReserVation Protocol," *ConneXions,* Volume 8, No. 8, August 1994.

[19] F. Flückiger, "Back to Basics: Networking requirements of audio and motion video," *ConneXions,* Volume 10, No. 1, January 1996.

[20] M. Handley & J. Crowcroft,, "The Internet Multimedia Conferencing Architecture," *ConneXions,* Volume 10, No. 6, June 1996.

[21] S. Deering, D. Estrin, D. Farinacci, V. Jacobson, C-G. Liu, L. Wei, "An Architecture for Wide Area Multicast Routing," ACM SIGCOMM 1994, London October 1994, ACM *Computer Communications Review,* Volume 24, No. 4, pp 126–135.

[22] S. Deering, "Multicast Routing in Internetworks and Extended LANs," ACM SIGCOMM 1988, August 1988, pp 55–64.

[23] H. Schulzrinne, S. Casner, R. Frederick & V. Jacobson, "RTP: A Transport Protocol for Real-Time Applications," Internet-Draft, Work In Progress, Late 1995.

[24] M. Handley, V. Jacobson, "SDP: Session Description Protocol," Internet-Draft, Work in Progress, November 1995.

[25] G. Chesson, "The Xpress Transfer Protocol (XTP)," *ConneXions,* Volume 5, No. 12, December 1991.

[26] V. Cerf, "Broadcasting and Multicasting," *ConneXions,* Volume 1, No. 1, May 1987.

[27] S. Deering, "IP Multicasting," *ConneXions,* Volume 5, No. 3, March 1991.

[28] S. Deering and S. Casner, "First IETF Internet Audiocast," *ConneXions,* Volume 6, No. 6, June 1992.

[29] S. Deering & D. Cheriton, "Multicast Routing in Datagram Internetworks and Extended LANs," ACM *Transactions on Computer Systems,* Volume 8, No. 2, May 1990.

MARKUS HOFMANN is a research assistant at the Institute of Telematics at the University of Karlsruhe. He received his Diploma degree in 1994 from the University of Karlsruhe. Currently, he is a member of the High Performance Networking Group and is working on protocol architectures for new generation networks. His research focuses on protocols for multimedia group communication. He can be reached as: `m.hofmann@ieee.org`

Announcement and Call for Participation

APRICOT, the *Asia Pacific Regional Internet Conference on Operational Technologies* will be held in Hong Kong January 27–31, 1997.

Background

Throughout Asia, Internet Service Providers, Backbone and Regional Networks, Web Hosting Facilities, Firewalls, and Private Intranets, are being installed at a staggering pace. The organizations responsible are under tremendous pressure to master the skills and policies necessary to operate and maintain these increasingly complex systems.

- APRICOT's mission is to satisfy this need for information. The conference consists of seminars, sessions, and forums with the goal of spreading and sharing the knowledge required to operate the increasingly complex Asia Pacific Internet topology.
- The First APRICOT was held in Singapore last January. It was attended by over 280 people from over 18 countries involved in delivering Internet Services.
- APRICOT's mission is to address the critical need to develop and advance the skills and understanding necessary to grow a robust Internet infrastructure in the Asia-Pacific region.
- APRICOT is not just another "pure-promotional" Internet conference. APRICOT is about bringing true subject matter experts together with those who can benefit from the information the most. The theme for this APRICOT will be "Managing the Growth of the Asia Pacific Internet."

Preliminary program

January 27 and 28: *Tutorials*—Introductory and Advanced Sessions Covering:

- Administering E-mail, News, and other Interactive Services
- The Domain Name System, Security, and General UNIX Systems Administration topics
- WWW Servers—Administration and Programming
- Basic TCP/IP Networking and Router Configuration
- Advanced Router topics, Routing Protocols and Interexchanges
- Content Control and Censoring Technologies

January 29 and 30: *Conference*—Tracks will include:

- Network Operations
- Applications and Services
- Internet Policy and Legal Issues
- Business of and on the Internet

January 31: *The Asia Pacific Network Information Center Meeting*

Sponsorship

Several levels of sponsorship are available. Contact the organizing committee for more information:

`apricot-oc@apricot.net`
c/o Mr. David Conrad
Telephone: +81 3 5467 7014
Facsimile: +81 3 5467 7015

More information

Watch `http://www.apricot.net/hk97` for upcoming final program and registration information.

Call for Papers

Baltzer Science Publishers in cooperation with ACM announce a Special Issue of the *Journal on Special Topics in Mobile Networking and Applications* (MONET) on *Mobile Networking in the Internet* with guest editors Charles E. Perkins, IBM T. J. Watson Research Center and David B. Johnson, Carnegie Mellon University.

Overview

The continued exponential growth of the Internet, coupled with powerful new technology for wireless computing, has created the need for extensive development of protocols and techniques for handling wireless nodes as they move about and change their point of network attachment. One major technological advance enabling mobile networking within the Internet has been Mobile IP, developed within the *Internet Engineering Task Force* (IETF), but this is only the beginning of the wave of changes needed to support nomadic Internet users. Problems introduced by mobility have been identified at every level of the network protocol stack, and many innovations are needed to enable the full potential of untethered nodes within the Internet.

Topics

This special issue will concentrate on the problems associated with mobile and wireless networking in the Internet, primarily at the network layer and above. A representative sampling of topics is provided below:

- Mobile IP
- Registration and location management
- Route optimization
- Interactions between geographic and network locality
- Mobile multicast protocols
- Mobile location of services and resources
- Transport layer (TCP, RTP) effects
- Multimedia and QoS support
- Internet protocols in a wireless ad hoc network
- Application adaptation to mobility and changing links
- Proxy architectures for Web and other services
- Internet security issues
- Analysis and simulation of mobile networking protocols
- Mobile node traffic analysis and simulation

Submission Guidelines

Authors should e-mail an electronic *PostScript* copy of their paper to one of the guest editors by *November 15, 1996*. Submissions should be limited to 20 double spaced pages, excluding figures, graphs, and illustrations. If e-mail submission is impossible, then six (6) copies of the paper should be sent by the due date to one of the guest editors.

Important dates

Manuscript due:	November 15, 1996
Acceptance notification:	January 31, 1997
Final manuscript due:	March 31, 1997

Guest Editors

Charles E. Perkins
Room H3-D34
IBM T.J. Watson Research Center
30 Saw Mill River Road
Hawthorne, NY 10532
Tel: +1-914-784-7350
Fax: +1-914-784-6205
E-mail: perk@watson.ibm.com

David B. Johnson
Computer Science Department
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213–3891
Tel: +1-412-268-7399
Fax: +1-412-268-5576
E-mail: dbj@cs.cmu.edu

This Call for Papers is also available on the Web at:

`http://www.cs.cmu.edu/~dbj/monet-cfp.html`

Future NetWorld+Interop Dates and Locations

NetWorld+Interop 96	Paris, France	October 8–11, 1996
NetWorld+Interop 96	London, England	Oct. 28–Nov. 1, 1996
NetWorld+Interop 96	Sydney, Australia	November 25–29, 1996
NetWorld+Interop 97	Singapore	April 7–11, 1997
NetWorld+Interop 97	Las Vegas, NV	May 5–9, 1997
NetWorld+Interop 97	Frankfurt, Germany	May 12–15, 1997
NetWorld+Interop 97	Tokyo, Japan	June 2–6, 1997
NetWorld+Interop 97	Atlanta, GA	October 6–10, 1997
NetWorld+Interop 97	Paris, France	October 20–23, 1997
NetWorld+Interop 97	London, England	October 27–30, 1997
NetWorld+Interop 97	Sydney, Australia	November 25–28, 1997

All dates are subject to change.

More information

Call 1-800-INTEROP or +1-415-578-6900 for more information. Or send e-mail to `info@interop.com` or fax to +1-415-525-0194. For the latest information about Interop DotCom and NetWorld+Interop as well as other SOFTBANK produced events, check our *Interop Online* home page at `http://www.interop.com`

NetWorld+Interop is produced by SOFTBANK Exposition and Conference Company, 303 Vintage Park Drive, Foster City, California 94404–1138, USA.

Write to *ConneXions*!

We'd love to hear your comments, suggestions and questions about anything you read in *ConneXions*. Our editorial address is given below. Use it for letters to the Editor, requests for the index of back issues, questions about particular articles etc.:

ConneXions—The Interoperability Report
303 Vintage Park Drive
Suite 201
Foster City
California 94404–1138
USA
Phone: +1 415-578-6900 or 1-800-INTEROP (Toll-free in the USA)
Fax: +1 415-525-0194
E-mail: `connexions@interop.com`
URL: `http://www.interop.com`

Subscription information

For questions about your subscription please call our customer service hotline: 1-800-575-5717 or +1 610-892-1959 outside the USA. This is the number for our subscription agency, Seybold Publications. Their fax number is +1 610-565-1858. The mailing address for subscription payments is: P.O. Box 976, Media, PA 19063–0976.

10.10

CONNEXIONS
303 Vintage Park Drive
Suite 201
Foster City, CA 94404-1138
Phone: 415-578-6900
FAX: 415-525-0194

ADDRESS CORRECTION
REQUESTED

FIRST CLASS MAIL
U.S. POSTAGE
PAID
SAN JOSE, CA
PERMIT NO. 1